Fakultät für Medizin der Technischen Universität München

Möglichkeiten und Limitationen automatischer Wirbelkörpersegmentierungen in Abhängigkeit von patientenspezifischen Faktoren und vorhandenen Trainingsdaten

Anna-Lena Mörtl

Vollständiger Abdruck der von der Fakultät für Medizin der Technischen Universität München zur Erlangung des akademischen Grades einer

Doktorin der Zahnheilkunde (Dr. med. dent.)

genehmigten Dissertation.

Vorsitz:	Prof. Dr. Herbert Deppe
Prüfer*innen der Dissertation:	1. apl. Prof. Dr. Jan St. Kirschke
	2. Priv.- Doz. Dr. Dr. Jochen Weitz

Die Dissertation wurde am 04.01.2023 bei der Technischen Universität München eingereicht und durch die Fakultät für Medizin am 18.07.2023 angenommen.

Inhaltsverzeichnis

Abkürzungsverzeichnis

AKRONYME

ABB	*Abbildung*
AG	*Anna-Lena Grau*
BWS	*Brustwirbelsäule*
CNN	Convolutional- neuronal-network
CT	*Computertomographie*
DXA	*Dual-Röntgen-Absorptiometrie*
HD	*Distanzmaß nach Hausdorff*
HWS	*Halswirbelsäule*
ID-RATE	*Identification-Rate*
JK	*Prof. Jan Kirschke*
KI	*Künstliche Intelligenz*
KV	*Kilovolt*
LWS	*Lendenwirbelsäule*
MRT	*Magnetresonanztomographie*
MW	*Mittelwert*
N	*Anzahl*
QCT	*Quantitative Computertomographie*
SD	*Standartabweichung*
WK	*Wirbelkörper*

1 Einleitung

Der demografische Wandel und die immer älter werdende Bevölkerung haben einen zentralen Einfluss auf die Medizin des 21. Jahrhunderts. Osteoporose gilt als weltweit häufigste Knochenerkrankung und beeinträchtigt mit steigendem Alter zunehmend die individuelle Lebensqualität betroffener Patienten.[1] Eine frühzeitige Diagnose der Osteoporose mit anschließender, patientenspezifischer Therapie kann die Progredienz der Erkrankung deutlich mindern und sich langfristig positiv auf Morbidität und Mortalität auswirken. Das gesetzte Ziel einer frühzeitigen Diagnosestellung im Frühstadium ist jedoch bislang zu selten der Fall.[2]

Chronische Schmerzen, Invalidität und psychische Beeinträchtigungen (Verminderung des Selbstwertgefühles, Depressionen etc.) werden mit fortschreitenden Krankheitsverlauf häufig beobachtet [3]. Mit jeder vorliegenden vertebralen Fraktur steigt das Risiko für notwendige Operationen, zusätzliche Therapiemaßnahmen und vermehrte Krankenhausaufenthalte an.[4] Die daraus resultierende ökonomische Last für die nationalen Kostenträger ist erheblich. So wurden die Gesamtkosten zur Behandlung der Osteoporose in der europäischen Union bereits im Jahr 2010 auf 39 Milliarden Euro geschätzt. [5]

Es stellt sich die Frage: „Was bedingt eine Unterdiagnostizierung von Osteoporose und wie kann dies zukünftig verhindert werden um die Patienten-Belastung zu senken, Outcomes zu verbessern, sowie Ressourcen zu sparen?“ Für die bildgebende Diagnostik liegt derzeit keine allgemeingültige Klassifikation und Definition der Wirbelkörperfrakturen vor. Demnach wird abhängig vom Betrachter eine qualitative Diagnose gestellt, deren standardisierte Reproduzierbarkeit nicht zwangsläufig ist und einer gewissen Variabilität unterliegt.[6] Wirbelkörperfrakturen weisen nur bei einem Drittel der betroffenen Patienten eine Schmerzsymptomatik auf, sodass bei einem hohen Prozentsatz von Betroffenen das Vorliegen einer entsprechenden Erkrankung vorerst unerkannt bleibt. Die Varianz der auftretenden Symptome ist derzeit noch nicht abschließend geklärt, wird jedoch bislang mit Anzahl und Schweregrad der Frakturen und deren Lokalisation in der Wirbelsäule in Verbindung gebracht.[7] Ebenso können Wirbelkörperdeformitäten vielfältige Ursachen haben. Hereditäre Entwicklungsstörungen, Traumata und maligne vertebrale Frakturen wie z.B. Wirbelkörpermetastasen müssen differentialdiagnostisch ausgeschlossen werden. Veränderungen der Wirbelsäule sind auch durch Morbus Scheuermann, Spondylitis, Morbus Paget, etc. möglich. Dadurch ist die Reliabilität der radiologischen Befundung vermindert und Fehldiagnosen sind nicht auszuschließen. [8,9]

Eine automatische Detektion und Segmentierung der Wirbelkörper ist aus klinischer Sicht

sinnvoll und für die Diagnostik von osteoporotischen Frakturen anzustreben. Belangvolle Informationen wie z.B. Volumen, Form und quantitative Biomarker der Wirbel können aus der Segmentierung von CT-, und MRT- Aufnahmen gewonnen werden. [10] Die herkömmliche manuelle Segmentierung stellt sich hierbei als zeitaufwändiges Unterfangen dar und gewährt ein gewisses Maß an Variabilität zwischen den verschiedenen Betrachtern. So wird mit zunehmender Größe der Datenmengen der Einsatz von automatisierten Algorithmen zwangsläufig an Relevanz gewinnen.[11] Bemerkenswerte Fortschritte lassen sich im Bereich der Bilderkennung durch die Anwendung von CNN‘s (Convolutional Neuronal Network) verzeichnen. Diese ermöglichen nach Prozessierung von Trainingsdaten die gezielte Extraktion von Bildmerkmalen. Die Verarbeitung von stetig wachsenden Datenmengen kann somit simplifiziert und beschleunigt werden. In Bezug auf die Wirbelsäulendiagnostik soll mit der Nutzung von Algorithmen, eine gezielte Abgrenzung einzelner Wirbelkörper zu umgebenden anatomischen Strukturen automatisch erfolgen. [12]

In dieser Dissertation wird die Leistungsfähigkeit des eigens entwickelten Deep-Learning-Algorithmus zur Segmentierung von Wirbelkörpern in CT-, und DXA-Aufnahmen untersucht. Zur Bewertung der Segmentierungspräzision des CNN erfolgt ein Vergleich mit manuell korrigierten Segmentierungsmasken, welche von zwei Korrektoren erstellt wurden. Die Interindividualität des manuellen Korrektors soll dabei berücksichtigt und potentielle Auswirkungen auf das Segmentierungsergebnis aufgezeigt werden. Zusätzlich werden diverse Einflussfaktoren, welche die Qualität der automatischen, vertebralen Segmentierung beeinträchtigen, umfassend diskutiert.

1.1 Definition und Formen der Segmentierung

In der medizinischen Bildverarbeitung ist die Segmentierung ein häufig eingesetztes Verfahren und per Definition die Unterteilung eines Bildes in örtlich zusammenhängende Bereiche. Basierend auf einem Algorithmus können somit für den Betrachter relevante, von unrelevanten Arealen abgegrenzt werden. In der Medizin ermöglicht die Segmentierung eine Differenzierung unterschiedlicher anatomischer Strukturen wie zum Beispiel Knochen, Blutgefäße, Gewebe etc. beziehungsweise eine Separation von gesundem zu pathologisch verändertem Gewebe. [13]

Je nach Komplexität des zu verarbeitenden Bildmaterials kommen diverse Segmentierungs-methoden zum Einsatz. Als einfachste Verfahrensweise hat sich die manuelle Segmentierung

etabliert, welche auf einer dreidimensionalen manuellen Beschriftung der gesuchten Region gründet. Mit dem Cursor wird das zu bearbeitende Areal umfahren, sodass die Segmentierung und Nachkontrolle in einer Serie aus Einzelschichten der CT- bzw. MRT- Aufnahmen erfolgt. [14] Dieses Vorgehen bedarf eines sehr hohen personellen- und zeitlichen Aufwands und verzeichnet bei der Anwendung eine geringe inter-, und intraindividuelle Reliabilität. [15] Alternativ können fortgeschrittene Algorithmen zum Einsatz kommen, welche selektiv anatomische Areale und Strukturen visualisieren. Dieser Prozess gründet auf der Definition von Eingangs-Parametern wie z.B. Größe, Dichte und Begrenzung. Aber auch die Verarbeitung von komplexen Zusammenhängen unter algorithmischer Beurteilung von Homogenitätskriterien, Schwellwerten oder der Verteilung von Gradienten ermöglicht die Detektion der gesuchten Strukturen. [16] Um diese Vorgänge zu forcieren, wurden sog. Deep-Learning-Algorithmen entwickelt. Diese können anhand einer iterativen Prozessierung von Trainingsdaten korrekte Aussagen über zukünftige Problemstellungen treffen. [17] Deep-Learning stellt eine spezifische Unterform der künstlichen Intelligenz (KI) dar. Die Konstruktion des lernbasierten Algorithmus leitet sich von einem biologisch neuronalen Netzwerk ab (CNN = convolutional neuronal network).[18] Unter abstrakter Betrachtung beinhaltet dieses eine Eingabeschicht („Input-Layer), mit welcher Daten aufgenommen werden. Über mehrere nicht linear verknüpften Zwischenschichten (Neurone bzw. „Hidden-layers) erfolgt eine Weiterverarbeitung der Informationen, bis diese über das „Output – Layer“ wiedergegeben werden. [19]

Im Bereich der medizinischen Bildverarbeitung nimmt die Anwendung von Deep-Learning-Algorithmen für diagnostische Zwecke („computer aided detection/ diagnosis – CAD) progredient zu. Die Abstraktion und 3D- Visualisierung gezielter Strukturen liefert in der Analyse von medizinischen Bilddateien relevante Rückschlüsse für Diagnose- und Therapieentscheidungen. [20] Eine konsequente Verbesserung von Effizienz, Präzision und Reliabilität der angewandten Verfahren soll generiert und Objektivität und Vergleichbarkeit geschaffen werden. [21] Ebenso kann trotz zunehmender Qualität, Größe und Menge der Bilddateien eine optimierte Prozessierung der Daten erfolgen. Die Bereitstellung von großen öffentlichen Datensätzen („Big Data“) zum Training und Vergleich der Algorithmen stellt hierfür eine wichtige Grundvoraussetzung dar.

1.2 Zielsetzung

Patienten, welche von osteoporotischen Frakturen betroffen sind, unterliegen einer erhöhten Mortalität. Nach vorrausgegangener proximaler Femurfraktur und bestehender Osteoporose liegt die Sterblichkeitsrate innerhalb von 12 Monaten nach dem Frakturereignis bei 20-30 Prozent. [22] Vor diesem Hintergrund wächst die Relevanz und Dringlichkeit der zeitnahen Etablierung eines computergestützten Detektions-Verfahrens für osteoporotisch veränderte Wirbelkörper. Das Arbeiten mit einem CNN fordert die Bereitstellung eines Datenpools aus manuell korrigierten Segmentierungsmasken, um den Deep-Learning Mechanismus fortlaufend zu trainieren. Dies beansprucht jedoch einen enormen Zeitaufwand, da jedes Voxel manuell in der 3D- Schichtung überprüft werden muss und pathologische Abweichungen, technische Störfaktoren etc. eine intensive Auseinandersetzung bedingen. Die Mitgestaltung des Lernprozesses des human-maschinellen Hybrid-Algorithmus zur Visualisierung von Wirbelkörpern im CT und MRT soll zu einer Vereinfachung der Osteporosediagnostik führen. Die Voraussetzungen dafür sind ein grundlegendes Verständnis für die Anatomie der Wirbelsäule und deren umgebende Strukturen, als auch Kenntnisse im Umgang mit manueller Segmentierung. Jedoch stellen sich hierbei vier Fragen die in der Dissertation erschlossen werden sollen:

1. Kann durch die manuelle Nachkorrektur der Segmentationsmasken unter Einbezug von medizinischem Wissen der algorithmische Outcome verbessert werden?
2. Welche Faktoren haben negative Auswirkungen auf die Segmentierung?
3. Besteht eine interindividuelle Abweichung beim Vergleich der manuellen Segmentierung zwischen zwei Korrektoren (einer Medizinstudentin und einem Professor für Radiologie) und kann medizinische Objektivität generiert werden?
4. Werden alle Abschnitte der Wirbelsäule vergleichbar präzise automatisch segmentiert?

Die zweite Fragstellung hinsichtlich der Einflussfaktoren stellt dabei eine bedeutsame zukunftsorientierte Komponente dar. Durch das Erkennen einzelner Faktoren mit negativen Auswirkungen auf die Segmentierungen können individuelle Lösungsstrategien entwickelt werden, um die Ergebnisse weiter zu optimieren. Ein langfristiges Ziel besteht im Übergang von einem Human-Hybrid-Algorithmus zu rein computergestützten Algorithmen und damit vollautomatischer Detektion osteoporotisch veränderterer Wirbelkörper. Eine Minimierung der Zufallsbefunde durch Erstdiagnostizierung im Frühstadium soll im Wesentlichen den Krankheitsverlauf positiv beeinflussen und die Belastung für Patienten verringern.

2 Material und Methoden

2.1 Versuchsaufbau und Ablauf

Die Arbeitsgruppe des Instituts für Radiologie – Abteilung für diagnostische und interventionelle Neuroradiologie des Klinikums rechts der Isar (Schwerpunkt Wirbelsäulendiagnostik) organisierte im Rahmen des Medical Image Computing and Computer Assisted Intervention (MICCAI 2019 Shenzen – China) die Large Scale Vertebrae Segmentation Challenge (VerSe Challenge 2019). Diese wurde mit einem Datenset von 160 Wirbelsäulen-CT-Scans und den dazugehörigen Bearbeitungs-tools öffentlich ausgeschrieben und von diversen internationalen Teams wahrgenommen.[23,24] Das Ziel der Challenge bestand in der Konstruktion und Weiterentwicklung eines lernbasierten Algorithmus zur Segmentierung von Wirbelkörpern auf Voxel-Ebene. Als derzeit größtes veröffentlichtes CT-Datenset der Wirbelsäule sollte ein uneingeschränktes Training der Algorithmen mit annotierten Daten ermöglicht werden, um das maschinelle Lernen zu fördern und die Detektion, Segmentierung und Frakturerkennung der Wirbelkörper zu automatisieren.[24,25]

Der eigens entwickelte Algorithmus benötigte zwei Arten von Anmerkungen um eine automatisierte Prozessierung zu ermöglichen. Darunter zählte die Etikettierung der Wirbelkörper auf Voxel-Ebene als auch die Annotation der dreidimensionalen Koordinatenpositionen der Wirbelschwerpunkte für alle 25 Wirbelkörper (C1-L6.) [24] Die Verarbeitung der Wirbelsäulen-Scans gliederte sich in einen dreistufigen Vorgang. Im ersten Schritt wurde anhand eins dreidimensional vollständig gefalteten neuronalen Netzwerks (CNN) die Wirbelsäule mit einer niedrig-aufgelösten Heatmap überlagert und somit die Detektion der Wirbel angestrebt. Unter Zuhilfenahme des Btrfly-Nets, welches sich sagittaler und koronaler Maximumintensitätsprojektionen des betroffenen Wirbelsäulenabschnittes bediente, erfolgte unter dem sog. Labeling eine Differenzierung der Wirbel.[26] Die Massenschwerpunkte der Wirbelkörper wurden in diesem Vorgang mit Zentroiden markiert. Die daraus resultierende Nummerierung der Wirbel stellte die Grundvoraussetzung für die Produktion einer 3D-Multilabel-Maske dar. Unter dem Aspekt der quantitativen Bildanalyse konnten durch die Verwendung eines optimierten U-Nets alle im CT-Scan abgebildeten Wirbelkörper segmentiert werden.[27] Auf Basis von zwei öffentlichen Datensätzen (Computational Spine Imaging und xVertSeg), welche innerhalb der Arbeitsgruppe analysiert und korrigiert wurden, erfolgte das erste Training des halbautomatischen Segmentierungs-Algorithmus. Die CT-Scans der Verse-Challenge bildeten die Grundlage für alle weiteren Trainingseinheiten. Im Zeitraum von Mai bis September 2019 wurden die automatisch generierten Zentroide und Segmentierungsmasken

des Datensatzes mittels Gebrauch der ITK-SNAP Software unter der Leitung von Prof. Dr. Kirschke überprüft und bei Bedarf manuell korrigiert. Die manuelle Segmentierung wurde von 2 Korrektoren (Erstkorrektur einer von vier Studenten: Alina Jacob, Anna-Lena Grau, Andreas Scharr, Mareike Kallweit; Zweitkorrektur einer von zwei Radiologen: Prof. Dr. Jan Kirschke, Maximilian T. Löffler) unabhängig voneinander durchgeführt und gemäß der anatomischen Morphologie in jeder der drei CT- Ebenen angepasst. Das Basis-CT stellte eine „nii.gz“-Datei dar, mit der identischer Ausrichtung und Einteilung wie die zu überlagernde vertebrale Segmentierungs-Schablone. Nach erfolgter Korrektur des algorithmischen Outcomes wurden die Dateien für weitere Trainingsebenen freigegeben. Dieses iterative Vorgehen sollte eine Erweiterung der Trainingsdaten, mit korrelierender Verbesserung des Algorithmus und Abnahme weitere notwendiger Korrekturen verzeichnen [24,25]

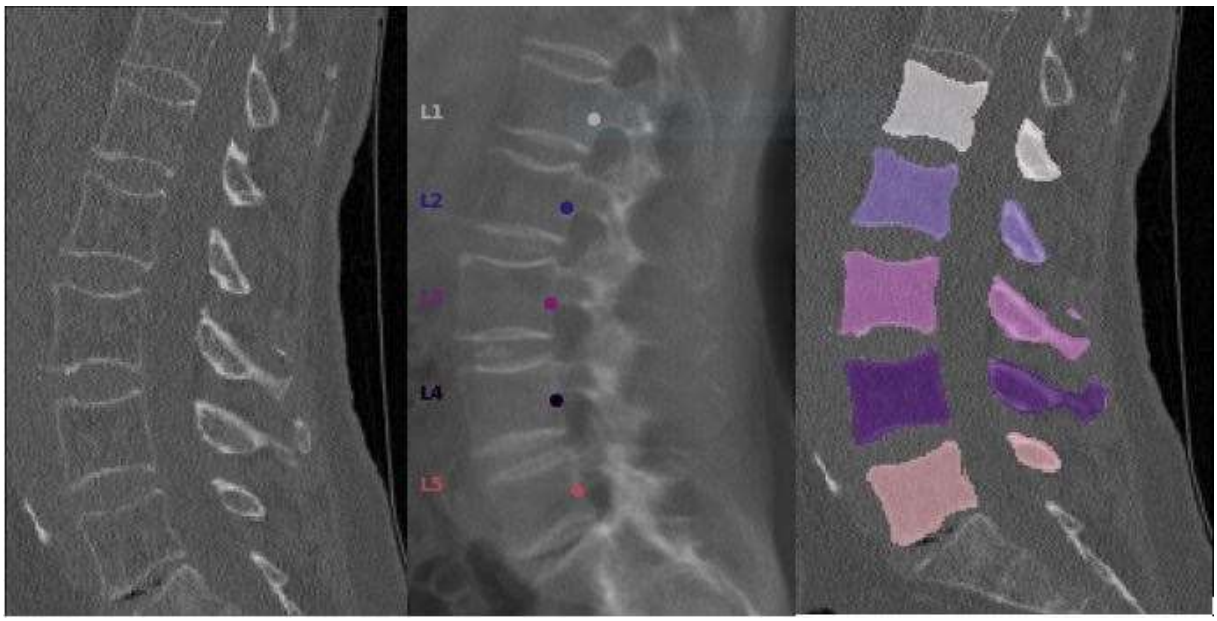

***Abbildung 1**: Sagittaler Ausschnitt der Lendenwirbelsäule mit Darstellung der algorithmischen Arbeitsschritte. Links Ausgangs-CT DX_12, in der Mitte L1-L5 mit Markierung der Zentroide und rechts Überlagerung der Wirbelkörper mit Segmentierungsmaske nach Zweitkorrektur.*

2.1.1 Patienten und Bildmaterial

Nach Antrag (27/19 S-SR) willigte die Ethikkommission des Klinikums rechts der Isar der TUM der retrospektiven Auswertung der Patientendateien ein. Die CT-Dateien, aus zwei retrospektiven Studien stammend, führten auf eines von zwei Einschlusskriterien zurück. Diese lauteten: das Vorhandensein einer nicht-kontrastmittelverstärkten Aufnahme der gesamten Wirbelsäule oder das Vorliegen einer lumbalen DXA-, als auch CT-Datei, aufgenommen

innerhalb eines Jahres mit abgebildeter Lendenwirbelsäule. Lediglich Aufnahmen mit einer kV-Einstellung von 120kvp und sagittalen Rekonstruktionen durch gefilterte Rückprojektion und spatialer Auflösung von ≤ 1mm in kraniokaudaler Richtung (Knochenkern) fanden Berücksichtigung. Unter diesen vorbehaltenen Aspekten konnte eine Kohorte von 454 Patienten selektiert werden, aus welchen 200 Aufnahmen zufällig ausgelost wurden. Die Patientenscans entstanden im Zeitraum von Januar 2013 bis November 2017 bei stationär behandelten Patienten. Das vorhandene Bildmaterial wurde sowohl für diagnostische als auch therapeutische Maßnahmen (z.B. Verifizierung von osteoporotischen Veränderungen der Wirbelsäule, akute bzw. chronische Rückenschmerzen, etc.) angefertigt. Mit einem von fünf Multidetektor CT- Scanner (Siemens Healthineers; Philips Brilliance 64, iCT 256, and IQon; Siemens Somatom Definition AS; Philips Medical Care) konnten die CT-Scans akquiriert werden. Eine Unterteilung in zwei bis drei Scan-Einheiten, korrelierend zum abgebildeten Wirbelsäulenabschnitt (Hals-, Brust-, Lendenwirbelsäule), erfolgte bei vereinzelten Patientenaufnahmen aufgrund des Scannerprotokolles. Im Laufe der Forschungsarbeit wurde die Anzahl der Bilddateien von 200 auf 160 bzw. 1725 Wirbel reduziert. Dies spiegelte schlussendlich eine Kohorte von 141 Patienten, mit einem Alter von ≥30 Jahren, bei vorhandener Absenz von Knochenmetastasierung, wieder. Die ausselektierten Datensätze konnten auf Grund pathologischer Abweichungen bzw. metallischer Störfaktoren nicht für das maschinelle Lernen des Algorithmus inkludiert werden. Die angewandten 3D-Scans wiesen Variationen in Bezug auf Anzahl der Wirbelkörper und der entsprechenden Wirbelsäulenabschnitte auf. (siehe Tab.1)

Tabelle 1: *Anzahl der segmentierten Wirbelkörper pro Wirbelsäulensegment im Rahmen der Verse-Challenge; Gliederung der Wirbelsäule in Hals-, Brust-, und Lendenwirbelsäule*

Wirbelsäulensegment	LWS	BWS	HWS
Anzahl x (WK)	621	884	220

Ct-Scans mit einer Bandbreite von 3-25 abgebildeten Wirbelkörpern wurden in das Datenset integriert. Im CT lediglich partiell dargestellte Wirbel fanden für die weitere Prozessierung keine Berücksichtigung. Ebenso stellte die Heterogenität des Datensatzes für das autonome Lernen des Algorithmus einen relevanten Faktor da. Daher wurden Bilddateien mit pathologischen Veränderungen, im Sinne von Osteophyten, osteoporotischen Frakturen, Wirbelkörper mit Zustand nach Trauma bzw. Spondylodese etc. explizit ausgewählt.[25]

2.1.2 Training n1-n9

Die Detektion, Etikettierung und Segmentierung der Wirbelkörper lassen sich zwar als separate Aufgaben des Algorithmus, jedoch nicht als voneinander unabhängige Abschnitte betrachten. Der Algorithmus arbeitet als „iBack-Framework“ und zielt in seinem Rahmenwerk auf die Kohärenz der differenzierten Aufgaben ab. Menschliche Interaktion ist in jedem der drei Aufgabenteile möglich und erlaubt Veränderungen am algorithmischen Outcome. So können die automatisch annotierten Zentroide nachträglich manuell repositioniert werden.[24] Um jedoch eine Automatisierung des Verfahrens zu ermöglichen, wird die Definition von Eingangsparameter benötigt. Durch die Annotation der Zentroide kann exemplarisch eine algorithmische Grundwahrheit für die Detektionsaufgabe manifestiert werden. [28]

Neben der Konstruktion des Algorithmus müssen jedoch auch kontinuierlich Trainingsdaten zur Verfügung gestellt werden. Das Training des Algorithmus im Rahmen der Verse Challenge vollzog sich mit 80 CT-Scans und jeweils dazugehöriger Segmentierungsmaske. Die Dateien wurden im Rahmen von 9 Trainingseinheiten (n1-n9) aufgeteilt und dem Algorithmus partiell zum Training freigegeben. Nach dem ersten Training (= n1) erfolgte die Begutachtung der produzierten Segmentierungsmasken, sodass bei Bedarf eine manuelle Nachkorrektur erfolgen konnte. Anschließend wurde die erste Trainingseinheit mit zusätzlichen CT-Scans ergänzt und dem Algorithmus erneut zum Training freigegeben (= n2). Dieses Prozedere vollzog sich insgesamt neun Mal, um mit jeder weiteren Trainingseinheit die Präzision der Segmentierung durch den Deep-Learning-Mechanismus des Algorithmus zu verbessern.

2.1.3 Release 1-4

Das Datenset von 160 Ct’s wurde in vier Release unterteilt. Release 1 und 2 repräsentierten das Trainingsset mit 80 Scans, Release 3 und 4 stellten Test-Phase 1 und Testphase 2 mit jeweils 40 Scans dar. Die Veranstalter der Challenge publizierten das vollständige Trainingsset inkludierend der Bilder, Segmentationsmasken und Zentroid-Anmerkungen in drei Phasen (16. Mai, 11. Juni, 17 Juli 2019) unter Mitbeteiligung der Teams. Nachfolgend wurden in Test-Phase 1 die identischen CT- Scans an alle Teilnehmer ausgehändigt und diese aufgefordert bis zum 23. August den Outcome ihres Algorithmus und einen Bericht über ihr spezifisches Vorgehen einzureichen. Die Daten der Testphase-2 sollte nicht für die konkurrierenden Teams zugänglich sein, um ein Training des Algorithmus mit den CT-Scans und somit eine Verfälschung des Segmentierungsergebnisses zu ermöglichen. Mit einer Befristung von zwei Wochen wurden

die Teilnehmer aufgefordert ihren Algorithmus in einem Docker-Container zu veröffentlichen, sodass dessen Leistungsfähigkeit anhand des versteckten Test-Datensatzes hinsichtlich der Bewertungskriterien überprüft werden konnte. Die Ermittlung des erfolgreichsten Algorithmus und des dazugehörigen Teams erfolgte im Anschluss. [24]

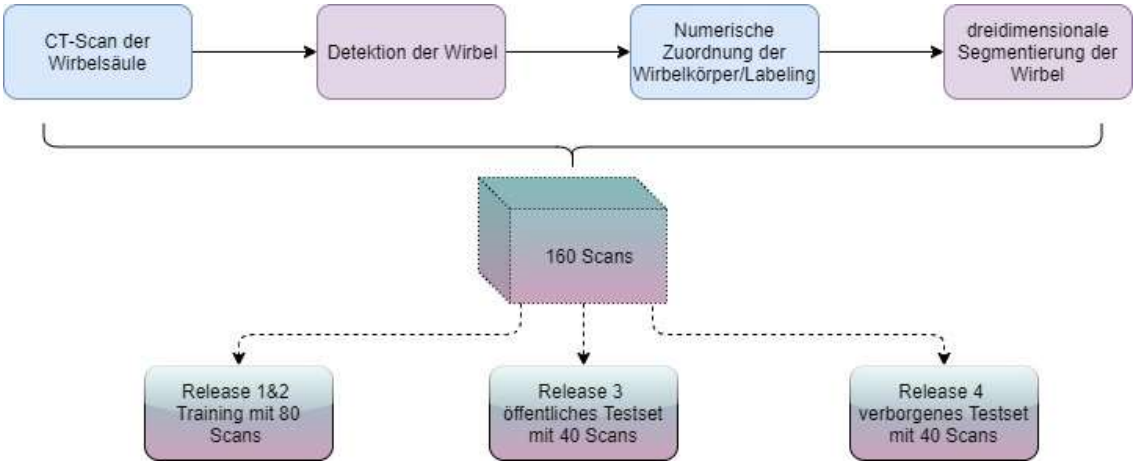

Abbildung 2: *Workflow der Verse-Challenge. Überblick über Verarbeitung der Scans und die Aufteilung des Datensatzes in unterschiedliche Release.*

2.1.4 Bewertungsmetrik der Challenge

Die Entwicklung eines hochfunktionalen vollautomatischen Algorithmus sowohl für die Etikettierungsaufgabe, als auch die Segmentierung der Wirbel war die Zentralaufgabe der Verse-Challenge. Anhand eines einheitlichen Bewertungsmaßstabs sollte der Algorithmus mit dem besten Labeling-, und Segmentierungsergebnis ermittelt werden. Von den Veranstaltern wurden zwei Bewertungsparameter je Aufgabe festgelegt. Der Dice–Koeffizient und der Hausdorff-Flächenabstand fanden bei der Segmentierungschallenge zur Qualitätsbestimmung Anwendung. Unter Berücksichtigung der beiden Metriken konnte die Segmentierungspräzision der Algorithmen beurteilt und durch deren Vergleich das beste Team eruiert werden.[24] Das Distanzmaß nach Hausdorff (HD) bezeichnet den lokalen Maximalabstand zwischen Segmentierungsmaske und Referenzvolumen in mm und quantifiziert mit steigendem Abstand eine Abnahme der Segmentierungsqualität. [20] Die Berechnung des Dice- Koeffizienten stellt sich dar als:

$$C_D = \frac{2\,|A \cap B|}{(|A| + |B|)}$$

Dabei sind A und B die Anzahl der Voxel welche pro Wirbelkörper zu den vergleichenden Segmentierungsmasken gehören. Die Menge der Voxel, welche sich in den Segmentierungen A und B überschneiden, wird durch $A \cap B$ bezeichnet. Mit steigender Anzahl der sich räumlich überlagernden Voxel nähert sich der Dice-Koeffizient dem Wert 1 an. Wird jedoch keine Übereinstimmung zwischen Voxelmenge A und B erzielt, ergibt sich daraus ein Dice-Koeffizient von 0. [29] Auch für die sich anschließenden Vergleiche unter 2.2.2. wird der Dice-Koeffizient als Bewertungsmaßstab herangezogen

Im Sinne der Etikettierungsaufgabe spiegelte die Identifizierungsrate (ID-Rate in Prozent %) das Verhältnis aller im Testsatz korrekt identifizierten Wirbelkörper wieder. Bei einem Abstand von weniger als 20mm zwischen der algorithmisch bestimmten 3D- Position des Wirbels und der wahrheitsgemäßen Wirbelposition erfolgte eine Inklusion in die Identifizierungsrate. Die örtliche Trennung der zu vergleichenden Wirbelpositionen wurde mit der zweiten Metrik - der Lokalisationsdistanz (in mm) bewertet.[26]

2.1.5 Ergebnisse der Challenge

Die Challenge und damit alle geforderten Aufgaben wurden nur vollständig von 7 Teams (anfänglich 18) erfüllt. Die Mannschaften Braun, AlibabaDamo, INIT und Huyujin wurden nicht allen Elementen des Wettbewerbs gerecht und fanden nur in denen von Ihnen erfüllten Aufgaben Berücksichtigung. Alle Teilnehmer wurden durch ein Punktesystem direkt miteinander verglichen. Bei besseren Ergebnissen einer Mannschaft gegenüber einer anderen wurde ein Punkt vergeben, sodass die Mannschaft mit den meisten Punkten als Sieger der jeweiligen Aufgabe hervorging. iFLYTEK erreichte gegenüber seinen 10 Kontrahenten das beste Dice-Score Ergebnis in Testphase 1 und erhielt somit 10 Punkte in dieser Rubrik. Die Ergebnisse der Testphase 2 wurden doppelt so stark gewichtet wie in Testphase 1, da durch die Unzugänglichkeit des Datensatzes eine spezifische Ausrichtung der Algorithmen und damit eine Beeinflussung der Resultate verhindert werden konnte. Die Etikettierung der Wirbel fand im Verhältnis 1:2 zur Segmentierungsaufgabe Berücksichtigung, da es auch als natürliches Produkt der Segmentierungschallenge hervorgehen konnte. Ebenso wurden die Metriken differenziert gewertet, d.h. D.mean und HD wurden im Vergleich zur id.rate und dem Dice-Score zur Hälfte berechnet. Christian_ Payer wurde als Sieger der Challenge ermittelt. Platz zwei belegte iFLYTEK und nlessmann erhob für sich den dritten Platz. Die Erstplatzierten Teams produzierten mit ihrem Herangehen an die Verse Challenge die führenden Ergebnisse.

In deren Verfahrensweise lassen sich jedoch wesentliche Unterschiede finden. So unterliegt im Team nlessmann die Prozessierung der Segmentierungssmaske einem differenzierten Ablauf. Die Wirbelkörper werden zuerst segmentiert und daraufhin erfolgt deren Labeling. In dieser Dissertation findet sich keine tiefergehende Analyse der Konzepte und Algorithmen anderer teilnehmender Teams. Jedoch lässt sich für zusätzliche Informationen auf den Report der Verse Challenge verweisen. (Verse: A vertebrae labelling and segmentation benchmark […]) [24]

2.2 Inter-Rater Segmentation Auto n4-n9 vs. Segmentation_AG vs. Final Ground Truth AG_jsk

Der erste Vergleich der Segmentierungsmasken bzgl. ihres Dice-Koeffizienten wird zwischen der automatisch generierten Segmentierungsmaske aus den Trainingseinheiten n4-n9, der ersten manuellen Korrektur „Segmentation_AG“ (korrigiert von Anna-Lena Grau - Studentin) und der zweiten Nachkorrektur „Final Ground Truth AG_jsk (korrigiert von Prof. Dr. Jan Kirschke - Radiologe) gezogen. J.K. weist dabei als Oberarzt und Experte für Neuroradiologie eine Berufserfahrung von 17 Jahren auf. Für den Vergleich liegen basierend auf der Verse Challenge 160 CT-Scans der Wirbelsäule mit zweifach unabhängig voneinander korrigierten Segmentierungsmasken vor. Grundlegend muss bei den zu vergleichenden Dateien jedoch sichergestellt sein, dass diese nicht in den kontinuierlichen Lernprozess (Training) des Algorithmus integriert wurden, da sonst eine fälschliche Verbesserung des Segmentierungsergebnisses in n4-n9 resultiert. Daraus folgend reduziert sich die Anzahl der herangezogenen Patientenfälle, welche von AG als auch JK korrigiert wurden von 88 auf insgesamt 41. Abhängig von der Anzahl korrekt segmentierter Voxel fand für die erste Korrektur eine Trainings-Segmentierung aus n4-n9 Verwendung. Um eine arbeits-, und zeitminimierende Nachbearbeitung zu ermöglichen, wurde in 34 Fällen für die erste Nachkorrektur eine Autosegmentierung aus n8 gewählt. Für 6 CT-Scans fiel die Wahl auf eine Segmentierungsmaske aus n9 und nur bei Scan ATL_97 aus n4aug. Lediglich vollständig abgebildete Wirbelkörper wurden segmentiert und für Vergleichszwecke herangezogen. Zwei der DXA CT-Scans bilden Wirbelsäulen- Abschnitte mit Fremdkörpermaterial ab (DX_223& DX_344). Für diese 41 Segmentierungs-Triplets bestehend aus Ursprungsscan mit Segmentierungssmaske, Nachkorrektur 1 &2 liegen pro Wirbel zwei Vergleiche vor:

1. Der Dice- Score zwischen „Segmentation Auto n4-n9“ & „Segmentation_AG“
2. Der Dice- Score zwischen „Ground Truth AG_jsk“ & „Segmentation_AG“

Beide Vergleiche zielen darauf ab die Kongruenz zwischen zwei Segmentierungsmasken zu untersuchen. Für jeden segmentierten Wirbelkörper der 41 Datensätze wird der Dice-Koeffizient, wie in Kapitel 2.1.4 beschrieben, berechnet. Ebenso erfolgt eine Darstellung des Mittelwerts, Median und der Standartabweichung pro WK. Der Dice-Koeffizient soll im Falle des ersten Experiments Qualitätsunterschiede zwischen der Autosegmentierung und manuellen Segmentierung beleuchten. So können mögliche Störfaktoren der computertechnischen Segmentierung dargelegt und patientenindividuelle Einflussparameter erörtert werden. Ebenso

ist die Autosegmentierung als zukünftiger Goldstandard zu diskutieren. Das zweite Experiment untersucht die interindividuellen Abweichungen der manuellen Segmentierung zweier Korrektoren. Durch die Betrachtung des Dice-Koeffizienten, resultierend aus Nachkorrektur 1&2 (Kap.3.1.1), wird die Deckungsgleichheit beider Segmentierungsmasken analysiert. Der Mittelwert des Dice-Koeffizienten pro WK verzeichnet Rückschlüsse für die Interrater-Reliabilität der manuellen Segmentierung. Ebenso erfolgt ein Vergleich der Dice-Werte zwischen den zervikalen-, thorakalen und lumbalen Wirbelsäulensegmenten. Diesbezügliche Schwankungen der Metrik und deren Ursache werden in der Dissertation untersucht.

2.3 Statistik

Qualitative Variablen werden mit ihrer Häufigkeitsverteilung dargestellt. Quantitative Variablen werden als Mittelwert, Standardabweichung (SD) und Median dargelegt. Alle statistischen Vergleiche wurden mit der Software Microsoft Excel (© Office Professional Plus 2019; Microsoft® Seattle, WA, USA) erhoben.

3 Ergebnisse

3.1 Analyse des Dice- Score

3.1.1 Vergleich des Dice-Score zw. Segmentation Auto n4-n9 vs. Segmentation_AG vs. Final Ground Truth AG_jsk

Nach Berechnung des Dice-Koeffizienten jedes Wirbelkörpers wurde für alle 41 Datensätze der Mean-Dice ermittelt. Als statistische Parameter liegen Mittelwert, Median, Range und Standartabweichung pro WK für beide Vergleiche vor. Die detaillierte Auflistung der erhobenen Werte findet sich im Anhang unter 6.1. und 6.2. Wie in Abbildung 3 dargestellt, lässt sich für die Erstkorrektur durch AG sowohl für den Abschnitt der HWS-, als auch BWS-Werte eine signifikante Abweichung des Dice-Koeffizienten zur Interrater-Kurve beobachten. Im Bereich der HWS- Segmentierung schwanken die Dice-Werte mit einer Range von 0,86-0,94. Noch größer zeigt sich die Spannweite der BWS- Segmentierungen mit einer Range von 0,84-0,97.

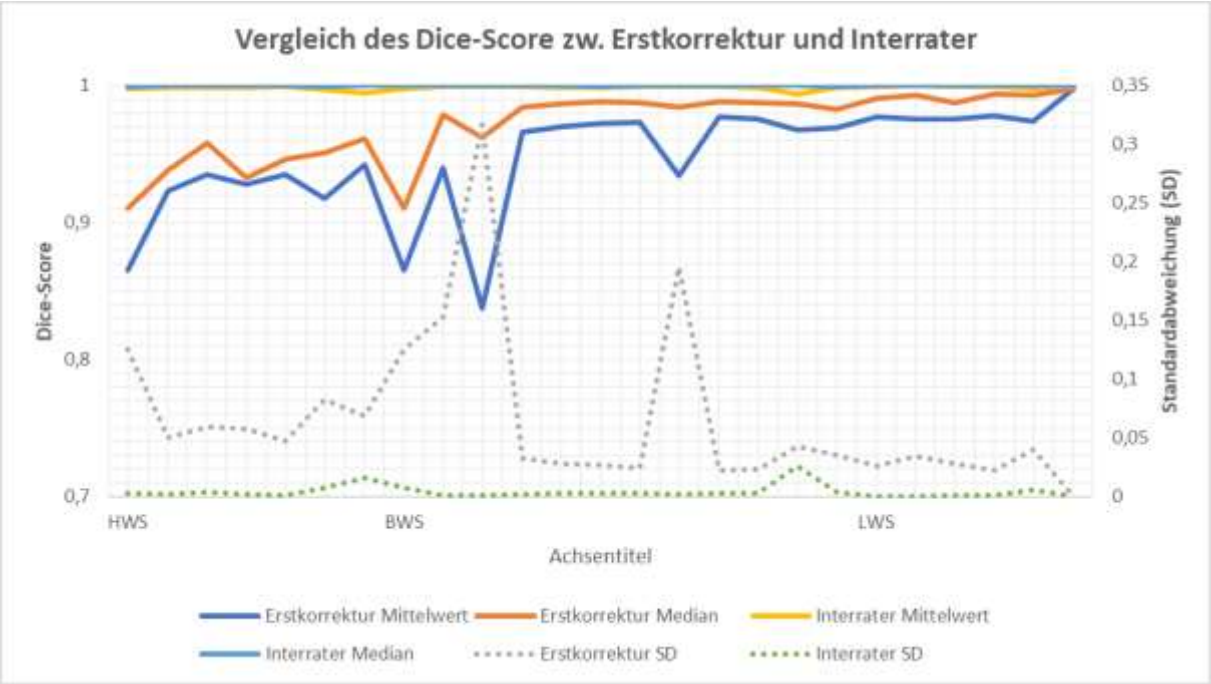

***Abbildung 3**: Vergleich des Dice- Score Mittelwert und Median pro WK in Bezug auf die Erstkorrektur (Segmentation Auto n4-n9 vs. Segmentation_AG) und die Zweitkorrektur/Interrater (Segmentation_AG vs. Final Ground Truth AG_jsk)*

Der Kurvenverlauf des MW der Erstkorrektur verdeutlicht im HWS-, und BWS- Segment eine verminderte Kongruenz zwischen Autosegmentierung und der manuellen Korrektur durch AG. Vereinzelte Wirbelkörper wie HW1, TH1 und TH3 zeigen unterdurchschnittliche Mittelwerte.

Diese sind bei allen drei Wirbelkörpern mit einer deutlich erhöhten Standartabweichung assoziiert. (Bsp. TH3, σ = 0,317). Im Bereich der LWS-Wirbelkörper lassen sich für die Erstkorrektur wesentlich konstantere Dice-Koeffizienten beobachten. Diese verzeichnen eine geringfügige Abweichung von bis zu 5% zum Interrater. Die gelbe Linie repräsentiert den Mittelwert der Dice-Koeffizienten nach Zweitkorrektur durch JK. Rückschließend veranschaulicht sie die Interrater- Reliabilität zwischen JK und AG welche über den gesamten Verlauf eine im Mittel deutlich erniedrigte SD von ± 0,004 zeigt.

***Tabelle 2**: T- Test bei abhängigen Stichproben zwischen den Dice-Score Mittelwerten pro WK für Erst und Zweitkorrektur*

Zweistichproben t-Test bei abhängigen Stichproben (Paarvergleichstest)		
	Erstkorrektur	***Interrater***
Mittelwert	0,947058882	0,998708422
Varianz	0,001649143	2,1612E-06
Beobachtungen	25	25
Pearson Korrelation	0,12779096	
Hypothetische Differenz der Mittelwerte	0	
Freiheitsgrade (df)	24	
t-Statistik	-6,384674426	
P(T<=t) einseitig	6,66757E-07	
Kritischer t-Wert bei einseitigem t-Test	1,71088208	
P(T<=t) zweiseitig	1,33351E-06	
Kritischer t-Wert bei zweiseitigem t-Test	2,063898562	

Eine statistisch signifikante Änderung der Dice-Score Mittelwerte je Wirbelkörper zwischen Erst und Zweitkorrektur lässt sich anhand Tabelle 2 belegen. Diese zeigt das Resultat eines durchgeführten T- Tests für die beiden abhängigen Stichproben (Erstkorrektur & Interrater). P(T<=t) einseitig vermisst mit 6,66757E-07 einen Wert, welcher deutlich kleiner als das festgelegte Signifikanzniveau von 0,05 ist. Daraus kann eine signifikante Verbesserung der Dice-Score-Mittelwerte durch manuelle Korrektur der Datensätze und Training des Algorithmus belegt werden. Die Effektgröße nach Cohens d_z zeigt für den durchgeführten T-Test mit 1,2768 einen großen Effekt.

Die Berechnung der Effektgröße ergibt sich wie folgt [30]:

$$d_z = \frac{T}{\sqrt{(df+1)}} = \frac{|-6{,}3847|}{\sqrt{25}} = 1{,}2768$$

3.1.2 Analyse des Dice- Score pro Wirbelsäulensegment

Bei einer Häufigkeitsverteilung der Wirbelkörper von HWS n=97, BWS-LWS n=304 und LWS n= 139 wird im zweiten Experiment die Deckungsgleichheit zwischen Autosegmentierung und Erstkorrektur in Korrelation zum betroffenen WS- Segment untersucht. (vgl. Abb.4) Alle Dice-Koeffizienten der Wirbelkörper, zugeordnet zum jeweiligen WS-Segment, bilden die drei Boxplots für HWS, BWS und LWS.

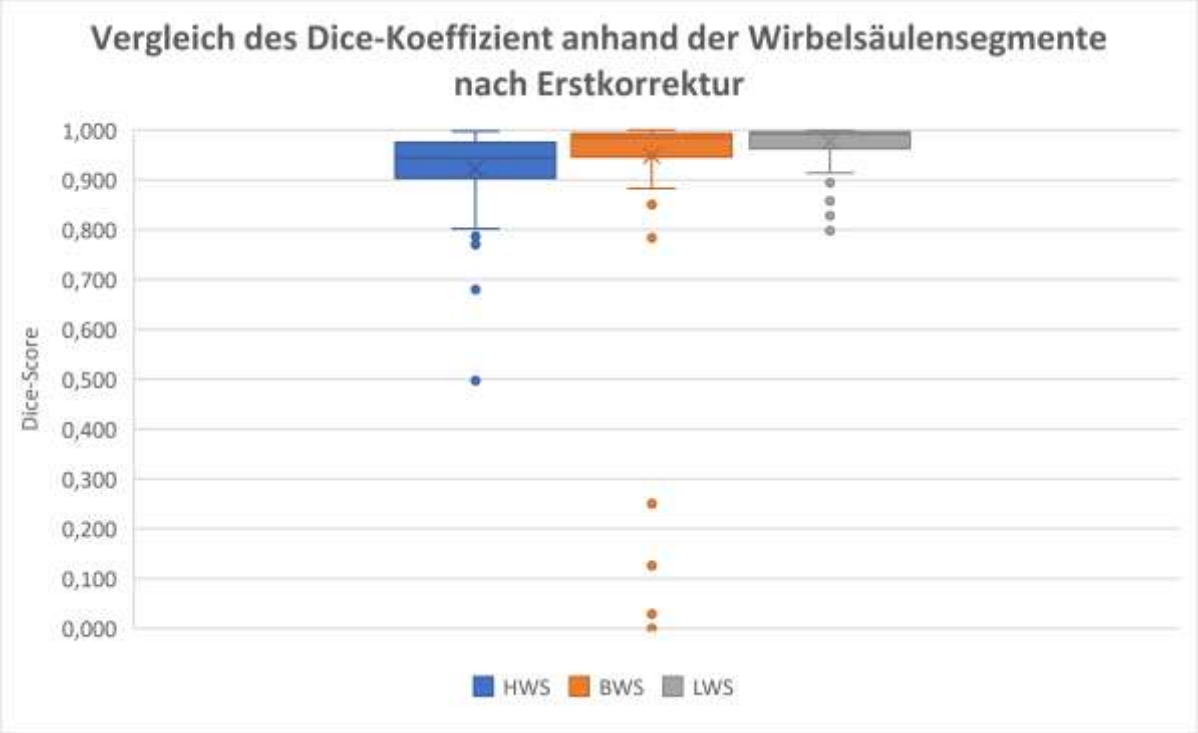

***Abbildung 4**: Einfacher Boxplott zum Vergleich des Mittelwerts des Dice-Koeffizienten bezogen auf das Wirbelsäulensegment; Die Dice-Score-Werte des Diagramms resultieren aus der Erstkorrektur des Datensatzes (Segmentation Auto n4-n9 vs. Segmentation_AG)*

Bezüglich des Vergleichs „Segmentation Auto n4-n9 vs. Segmentation _AG“ liegt ein Mittelwert von 0,923 (σ= 0,07) für die HWS-WK, von 0,952 (σ=0,13) für die BWS-LWS-WK und 0,977 (σ=0,03) für die LWS-WK vor. Dementsprechend kann in der Grafik gezeigt werden, dass die größte Übereinstimmung zwischen der Autosegmentierung und der manuellen

Segmentierung für die Werte der LWS Fraktion zu finden ist. Das Ergebnis der Halswirbelsäule verzeichnet mit einer Range von 0,497-0,997 die größte Abweichung und lässt somit auf eine verminderte Segmentierungspräzision des CNN mit einhergehend erhöhter Fehleranfälligkeit schließen. Ebenso werden in Abb. 5 die Mittelwerte der Mean-Dice-Scores pro Wirbelkörpersegment für den Vergleich aus manueller Erst-, und Zweitkorrektur dargestellt. Der Boxplot zeigt für alle drei Wirbelsäulenabschnitte vergleichbare Werte. Bei Betrachtung der Standartabweichung der Mean-Dice-Scores zeigt sich sowohl für HWS, BWS-LWS und LWS ein Wert < 0,005. Dies lässt darauf zurückführen, dass eine hohe Deckungsgleichheit aus der Segmentierung von AG und JK zu Grunde liegt.

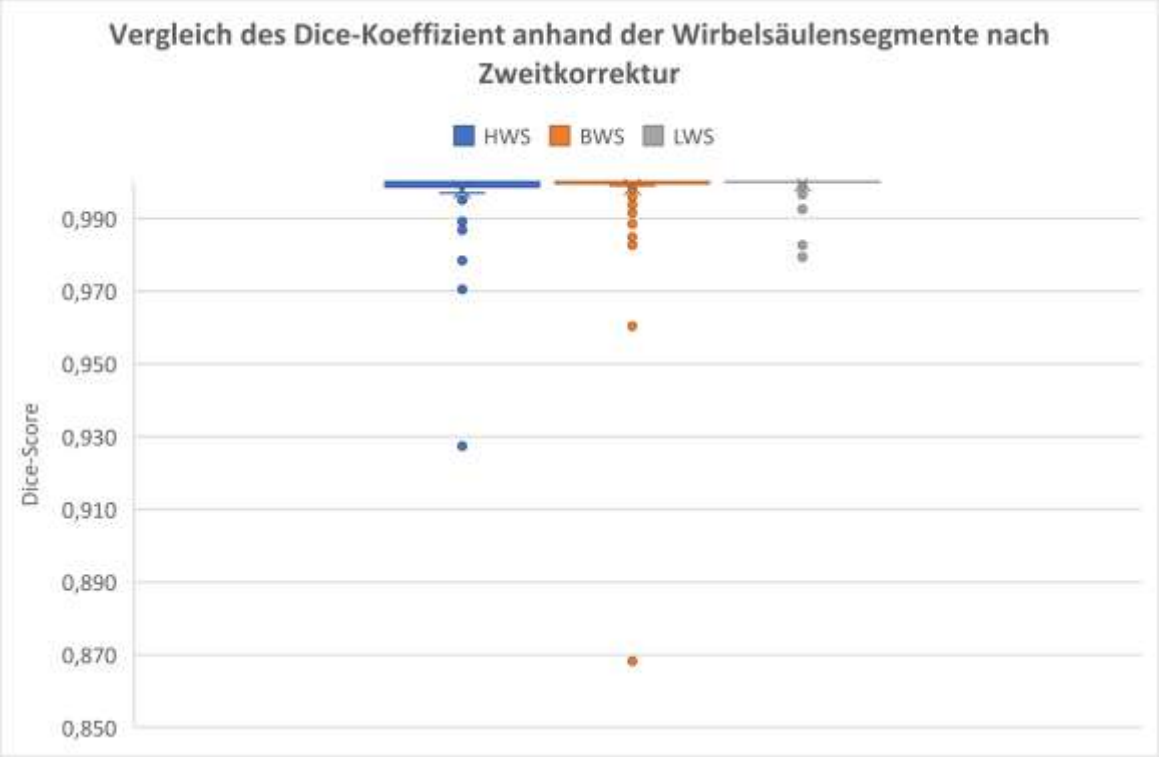

***Abbildung 5**: Einfacher Boxplot zum Vergleich des Mittelwertes der Dice-Koeffizienten bezogen auf das Wirbelsäulensegment; Die Dice-Score-Werte des Diagramms resultieren aus dem Interrater des Datensatzes (Segmentation_AG vs. Final-Ground-Truth_AG_jsk)*

4 Diskussion

4.1 Zusammenfassung der Ergebnisse

Die Dissertation ermöglicht einen grundlegenden Einblick in die derzeitigen Probleme der automatisch generierten Segmentierung der Wirbelsäule. Ein Algorithmus-gestütztes Verfahren kann dank multipler Trainingsebenen und zunehmender Anzahl an Trainingsdaten das Segmentierungsergebnis schrittweise verbessern, benötigt jedoch weiterhin humane Interaktionen und Korrekturen. Die vorliegenden Ergebnisse manifestieren die Notwendigkeit dieser manuellen Segmentierung, um Störfaktoren wie Rauschen, Artefakte, schlechter Kontrast sowie anatomische Abweichungen und degenerative Veränderungen zu berücksichtigen.

4.2 Abweichungen des Dice-Score zwischen Autosegmentierung und Erstkorrektur

Mit dem ersten Experiment - dem Vergleich zwischen Autosegmentierung und manueller Segmentierung lassen sich zwei grundsätzliche Fragestellungen beantworten. Diese lauten:

1. Welche Faktoren haben negative Auswirkungen auf das Segmentierungsergebnis?
2. Welche Patientenfälle stellen derzeit für den Algorithmus ein potentielles Problem dar?

Die Kurve aus Erstkorrektur durch AG und der Autosegmentierung in Abb. 3 unterliegt starken Schwankungen. Die vorliegend niedrigen Dice-Koeffizienten, explizit im Bereich der HWS und BWS Wirbelkörper repräsentieren eine verminderte Kongruenz der automatisch generierten Segmentierungsmaske mit der manueller Erstkorrektur. Dabei lassen sich anatomische Variationen und Abweichungen im CT feststellen, welche mit einer gehäuften Fehleranfälligkeit des Algorithmus einhergehen und eine mangelnde Segmentierungspräzision bei gezielten Datensätzen veranlassen. Diese lauten: Segmentierung osteoporotisch veränderter Wirbelkörper, Segmentierung von Osteophyten, Abgrenzung und Segmentierung der Wirbelkörper unter starker degenerativer Veränderung und Segmentierung der Wirbelkörper bei Zustand nach Trauma oder Spondylodese.

Eine fehlerhafte automatische Segmentierung tritt vermehrt an osteoporotisch veränderten Wirbelkörpern auf. Exemplarischer Weise wird dies anhand der Trainingseinheit Verse ID13 erörtert. Wie in Abb. 8 erkenntlich, kommt es durch den Deckplatteneinbruch an L1 zu einer Übersegmentierung durch den Wirbelkörper TH 12. Die Wedge Fraktur an L1 wird als solche nicht vom Algorithmus berücksichtigt und führt zu einer undefinierten Segmentierung des WK. L2 weist eine bikonkave osteoporotische Fraktur auf. Die Deformität des Wirbelkörpers ist jedoch im Vergleich zu L1 schwächer ausgeprägt und fordert nur eine geringe manuelle Korrektur. Dies spiegelt sich auch durch einen im Vergleich mit den restlichen Wirbelkörpern deutlich erniedrigten Dice Score von 0,894 für L1 nach Erstkorrektur wieder (s. Tabelle 3).

***Tabelle 3**: Ausschnitt der Dice- Scores aus dem Vergleich der 1. Korrektur und der automatisch generierten Segmentierungsmaske; Werte für Wirbelkörper TH11-L5 der Verse ID13; Niedrigster Score an L1(rot) gekennzeichnet;*

Wirbelkörper	*TH11*	*TH12*	*L1*	*L2*	*L3*	*L4*	*L5*
Dice-Score	0,946	0,944	**0,894**	0,942	0,956	0,964	0,962

Die geringe Anzahl an osteoporotisch veränderten Wirbelkörpern im Datenset, impliziert für den CNN eine Erweiterung und Spezifizierung der Trainingsdaten. Zu Gunsten des Lerneffektes und zur Verminderung der Fehleranfälligkeit des Algorithmus werden weitere Trainingseinheiten notwendig. Das Einbeziehen von Patientenfällen mit osteoporotischen Frakturen und zunehmender pathophysiologischen Formvariation der WK, soll die gewünschte Verbesserung der CNN- Ergebnisse ermöglichen. (vgl. Abb. 6)

Das gleiche Vorgehen ist für Patientenfälle, in welchen Wirbelfrakturen auf Grund von metabolischen, degenerativen oder infektiösen Knochenveränderungen wie z.B. Osteomalazie, Morbus Gaucher etc. vorliegen, indiziert. Auch der Zustand nach einem Trauma kann das Auftreten von degenerativer Veränderung mit ausgeprägter Formvariation der WK begründen.[8] In diesem Fall ist eine Stabilisierung der Wirbelsäule mittels Spondylodese möglich.[31] Die Metallimplantate, welche zur Versteifung der Wirbelsäule eingebracht werden, können im CT Aufhärtungs-, und Photon-Starvation-Artefakte verursachen.[32] Die eingeschränkte Visualisierung der anatomischen Strukturen bei zusätzlich verminderter Bildqualität bedingt eine Beeinträchtigung der Autosegmentierung. Verse ID93 stellt einen Datensatz bei vorliegender Spondylodese dar (vgl. Abb.7). Die WK L2, L3 und TH11 weisen im Bereich der Implantate eine lückenhafte Segmentierung auf. Dies wird durch den

degenerativen Zustand der Wirbelkörper unterstützt, wodurch eine klare anatomische Abgrenzung zu den umgebenden Strukturen erschwert ist.

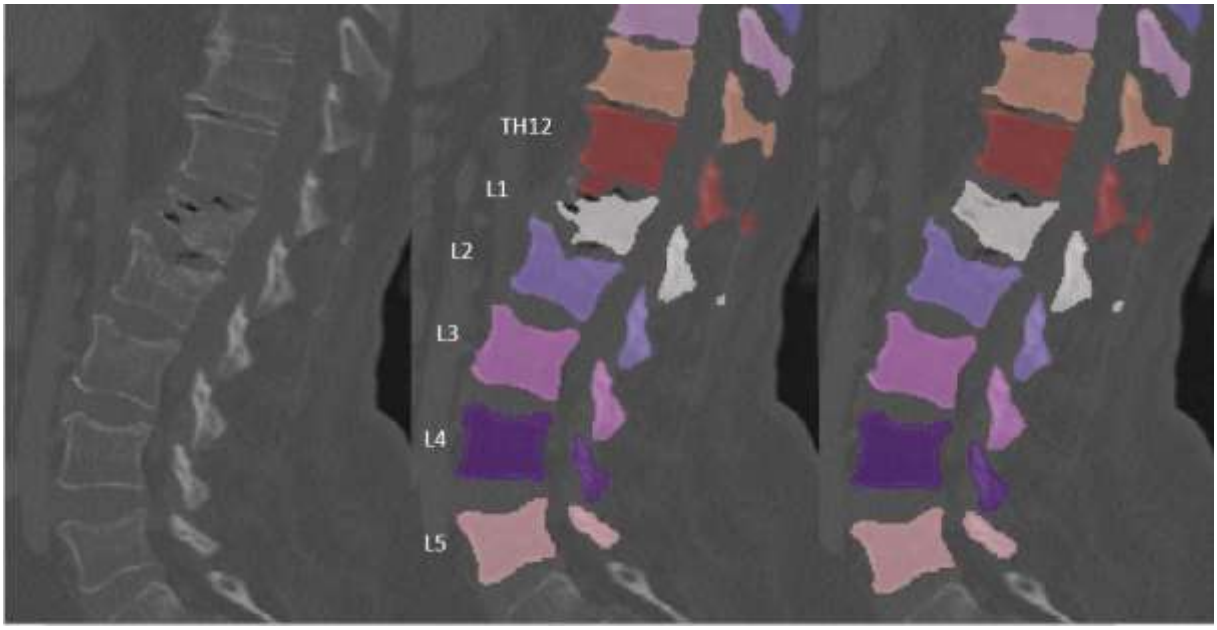

Abbildung 6: *Verse ID13 mit osteoporotischer Fraktur an L1 und L2; Ausgangs- CT (links), CT nach automatischer Segmentierung mit überlagerter Segmentierungsmaske; Fehlerhafte Segmentierung an TH12, L1 und L2(mitte); Korrigierte manuelle Segmentierung (rechts)*

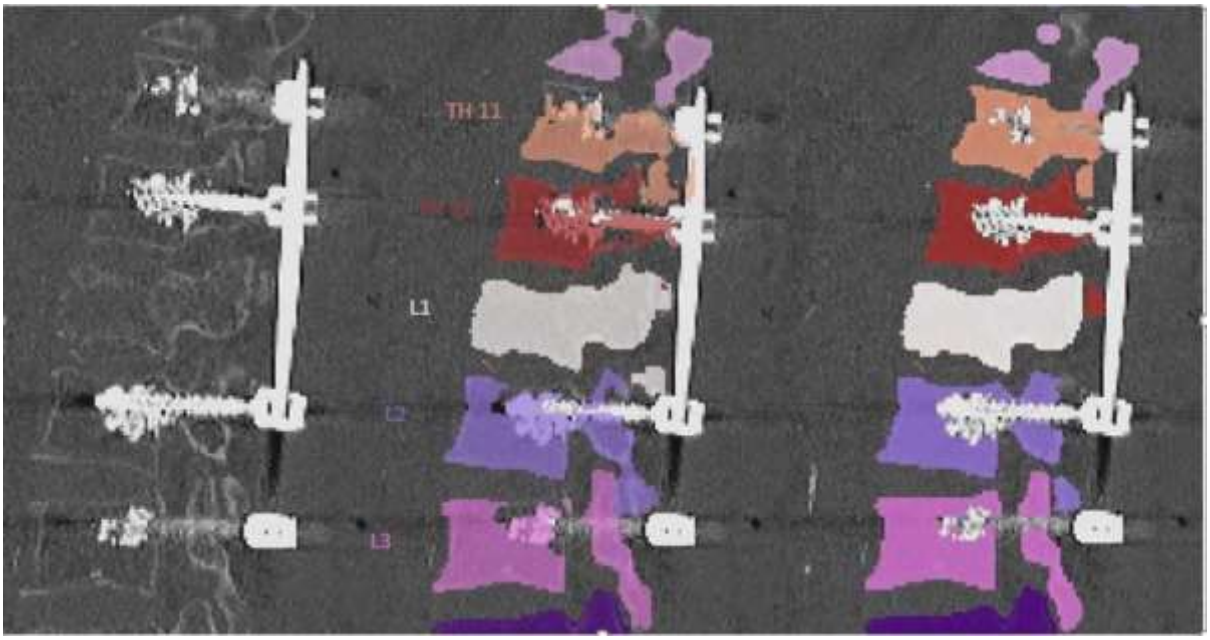

Abbildung 7: *Sagittale Darstellung der Spondylodese an den WK TH 11,12 & L2, L3 der Verse ID 93; Ausgang-CT mit schwacher Kontrastierung (links); Darstellung der automatisch generierten Segmentierungsmaske mit lückenhafter Segmentierung im Bereich der Deckplatten (Mitte); Manuell korrigierte Segmentationsmaske (Rechts);*

Zur Stabilisierung der Implantate wird in der Operation Actifuse, ein osteostimulatives Knochenersatzmaterial, im Bereich der Metallstäbe eingebracht. Dieses weist eine zum Knochen vergleichbare Dichte auf, wodurch die Differenzierung zwischen dem Knochenersatzmaterial und der natürlichen Hartsubstanz durch den Algorithmus nur schwer umgesetzt werden kann. Eine automatische Segmentierung des Actifuse kann resultieren. Die schwarzen Areale, welche an die Metallimplantate angrenzen, repräsentieren den Bereich des Artefakts. Auf Grund der fehlenden Bildinformation ist eine potentielle Zuordnung zum Wirbelkörper nur bedingt möglich. (vgl. Abb. 8) Die Patientenfälle bei vorliegender Spondylodese stellen diesbezüglich auch für den Korrektor eine wesentliche Herausforderung dar und beanspruchen eine aufwendige Nachbearbeitung. Diesbezüglich wurden lediglich zwei CT-Scans mit vorhandener Versteifung der Wirbelsäule für den Vergleich herangezogen. Jedoch kann nur durch gezieltes Training des CNN unter progredienter Diversität des Bildmaterials die Autosegmentierung verbessert werden.

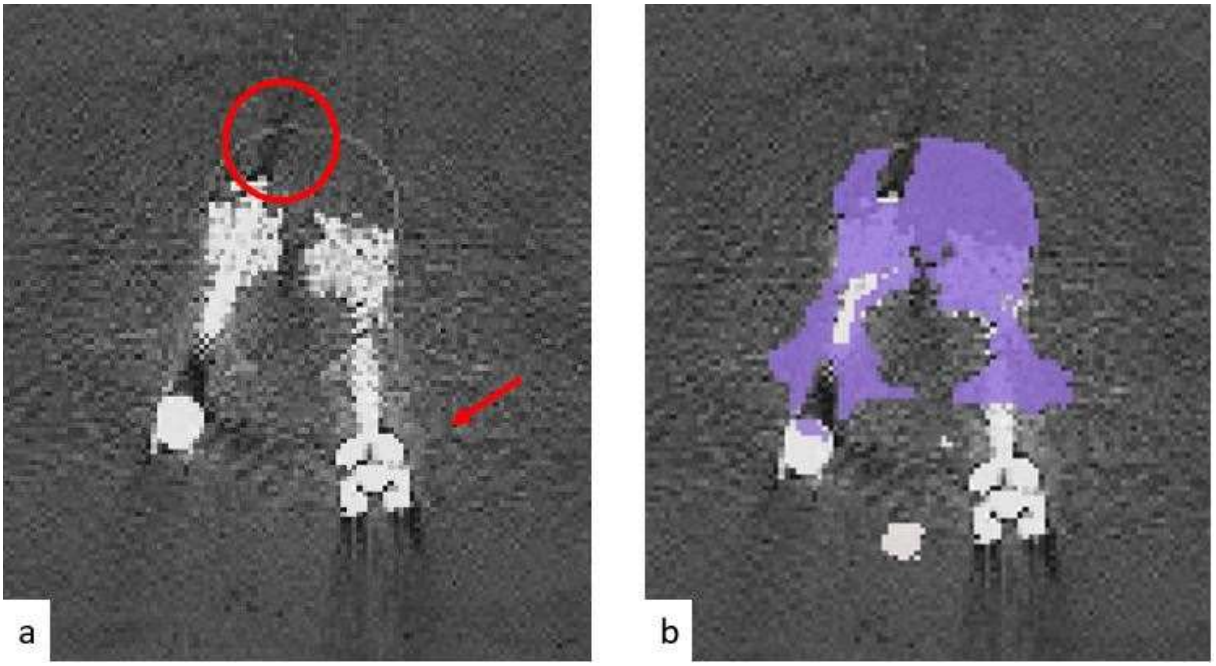

Abbildung 8: *Axiale Darstellung der Spondylodese an den WK L2 der Verse ID 93; Ausgang-CT mit schwacher Kontrastierung und Metallimplantat; Abbildung des Artefakts (Kreis) und des Knochenersatzmaterials (Pfeil) (links); Darstellung der automatisch generierten Segmentierungsmaske mit lückenhafter Segmentierung im Bereich der Artefakte und Überlagerung am Metallimplantat (rechts);*

Eine erhöhte Fehleranfälligkeit der Autosegmentierung zeigt sich auch bei der Segmentierung von Osteophyten. Bei diesem handelt es sich um einen knöchernen Auswuchs, welcher mit Faserknorpel bedeckt ist.[33] Das Erscheinungsbild der Osteophyten lässt sich häufig an den

antero-lateralen Rändern der Wirbelkörper betrachten.[34] Die grazile Form der Osteophyten wird jedoch nur bedingt vom Algorithmus automatisch erfasst, woraus partiell unsegmentierte Strukturen resultieren. In der 1&2 Nachkorrektur wurde die Segmentierung der Osteophyten explizit berücksichtigt, was die verminderte Kongruenz zwischen Auto-, und manueller Segmentierung mitbegründet. Die Abb. 9 des Datensatzes Verse ID80 zeigt einen thorakalen Wirbelsäulenabschnitt in sagittaler Darstellung mit ausgeprägten osteophytären Auswüchsen. Diese sind besonders an den Wirbelkörpern TH4-TH7 zu beobachten. Bei Betrachtung des Datensatzes mit zunehmender Trainingsebene n1-n9 lässt sich jedoch an den Osteophyten eine deutliche Steigerung der Segmentierungspräzision feststellen, was auf den Deep-Learning-Effekt des Algorithmus zurückgeführt werden kann.

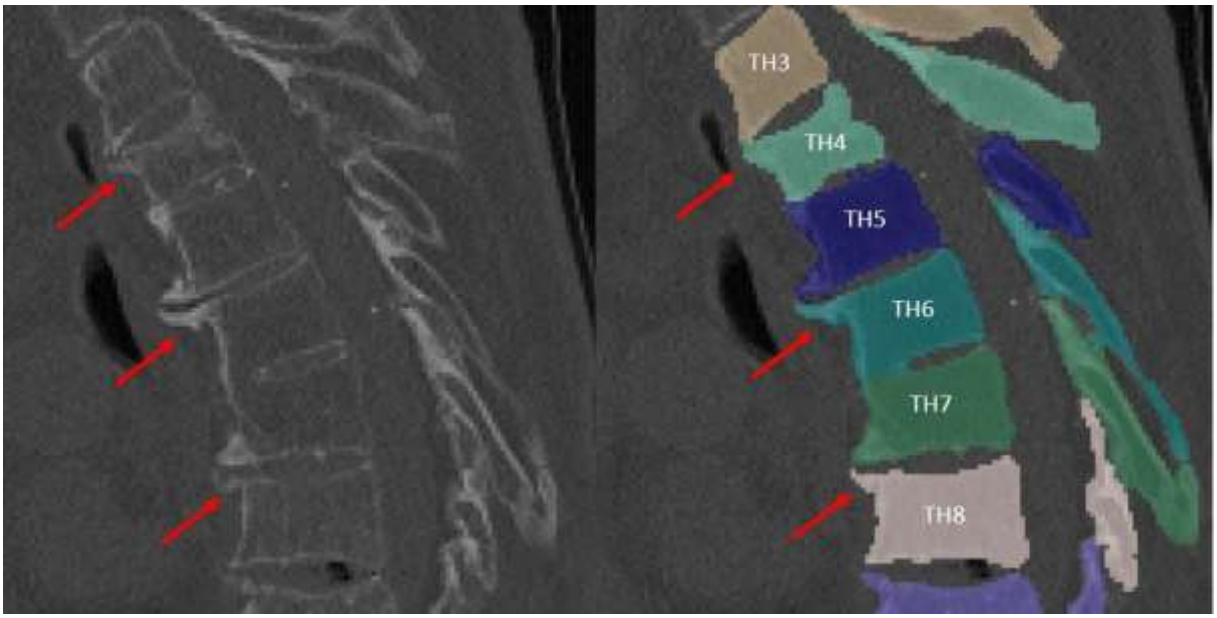

Abbildung 9: *Sagittale Darstellung der WK TH3-TH8 der Verse ID80; Ausgang-CT, Pfeile demonstrieren Osteophyten; (links) Manuell korrigierte Segmentierungsmaske; die Osteophyten wurden explizit berücksichtigt (Rechts);*

Grundsätzlich lässt sich sagen, dass die manuelle Nachkorrektur der Datensätze ein sehr zeitaufwendiger Arbeitsschritt ist. Bei vorliegend verminderter Segmentierungspräzision des CNN wurde ein Zeitaufwand von 1-6 Stunden pro Patientefall notwendig. Dieser ergab sich neben dem Einarbeiten in die Software-Koordination, der Auseinandersetzung mit vorliegenden Segmentierungstools, der Interpretation der WK- Anatomie, ebenso aus der Überarbeitung der Segmentierungsmaske für jeden Wirbelkörper (Größe pro WK mit 103 Voxel). Auch spielte der Deep-Learning-Mechanismus des Algorithmus und die Ausprägung

der anatomischen Variation der Wirbelkörper eine wesentliche Rolle. In den anfänglichen Trainingsstadien n1-n3 fielen korrelierend zur Fehleranfälligkeit des CNN vermehrt manuelle Korrekturen an. Mit zunehmender Anzahl der Trainingsebenen konnte jedoch die Präzision der automatischen Segmentierung verbessert werden, sodass sich auch der Aufwand an manueller Nachbearbeitung und investierter Zeit reduzierte. (vgl. Abb. 10) Eine kontinuierliche Weiterentwicklung des Algorithmus kann durch den Einsatz größerer Datenmengen und multifaktoriell variierten Patienteninformationen unterstützt werden.

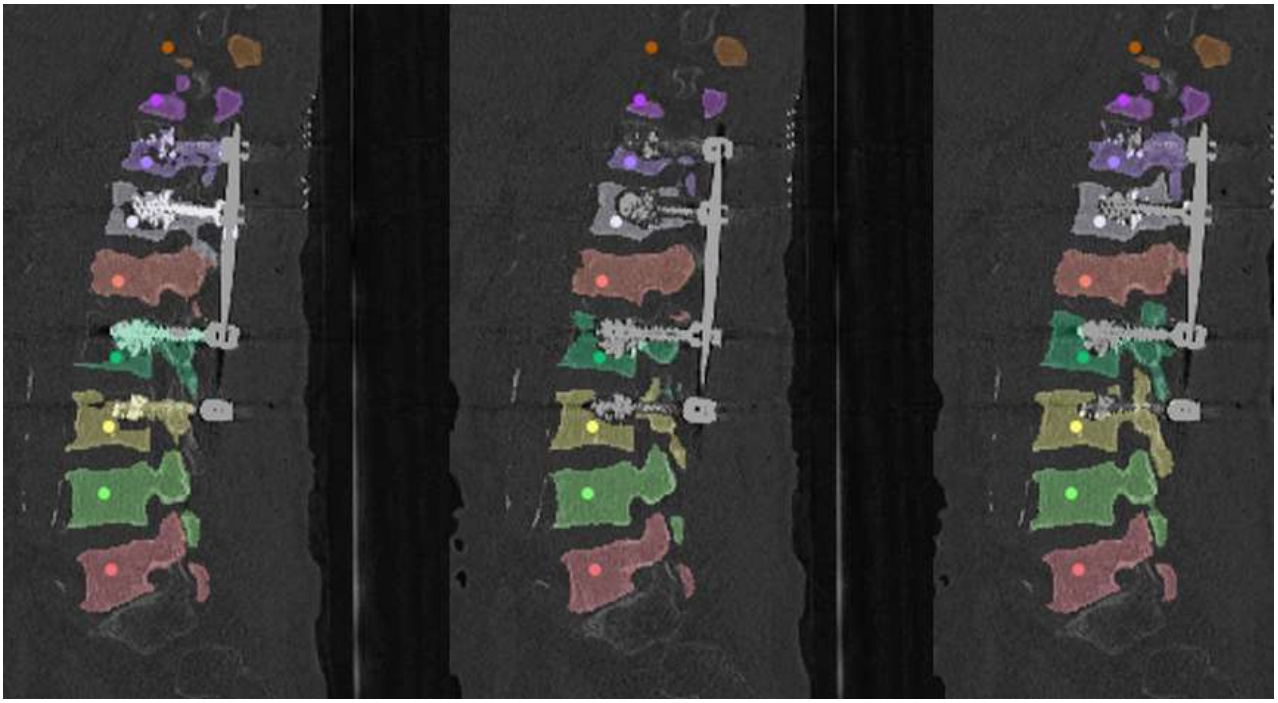

***Abbildung 10**: Datensatz Verse ID93 in unterschiedlichen Trainingsstadien N5(links), n6(Mitte), n8(rechts); Die Wirbelkörper wurden mit der mehrfarbiger Segmentationsmaske überlagert und die Zentroide als zentraler Punkt des WK dargestellt. Eine wesentliche Verbesserung des automatischen Segmentationsergebnis ist mit fortgeschrittenem Training ersichtlich (von links nach rechts). Besonders an L2 und L3 wird die Segmentationsmaske von links nach rechts zunehmend vervollständigt*

4.3 Abweichungen des Dice-Score zwischen manueller Erst-, und Zweitkorrektur

Die Interrater-Kurve in Abb.3 zeigt anhand des konstanten Verlaufs eine geringe interindividuelle Abweichung der manuellen Segmentierung zwischen Erst und Zweitkorrektur (AG & JK) auf. Durch die große Deckungsgleichheit der segmentierten Voxel wird ein Dice-Koeffizient nahe 1 bei jedem Datensatz erreicht. Die Konturen der Wirbel wurden mit präziser Abgrenzung zu umgebenen Strukturen dargestellt. Die abweichenden radiologischen Erfahrungswerte zwischen AG (Medizinstudentin) und JK (Prof. für Neuroradiologie) zeigen im Experiment keinen wesentlichen Einfluss auf die Interrater-Reliabilität, da auf eine Absprache grundlegender Segmentierungskriterien großer Wert gelegt und von Seiten JK anatomisch zu berücksichtigende Areale besprochen wurden.

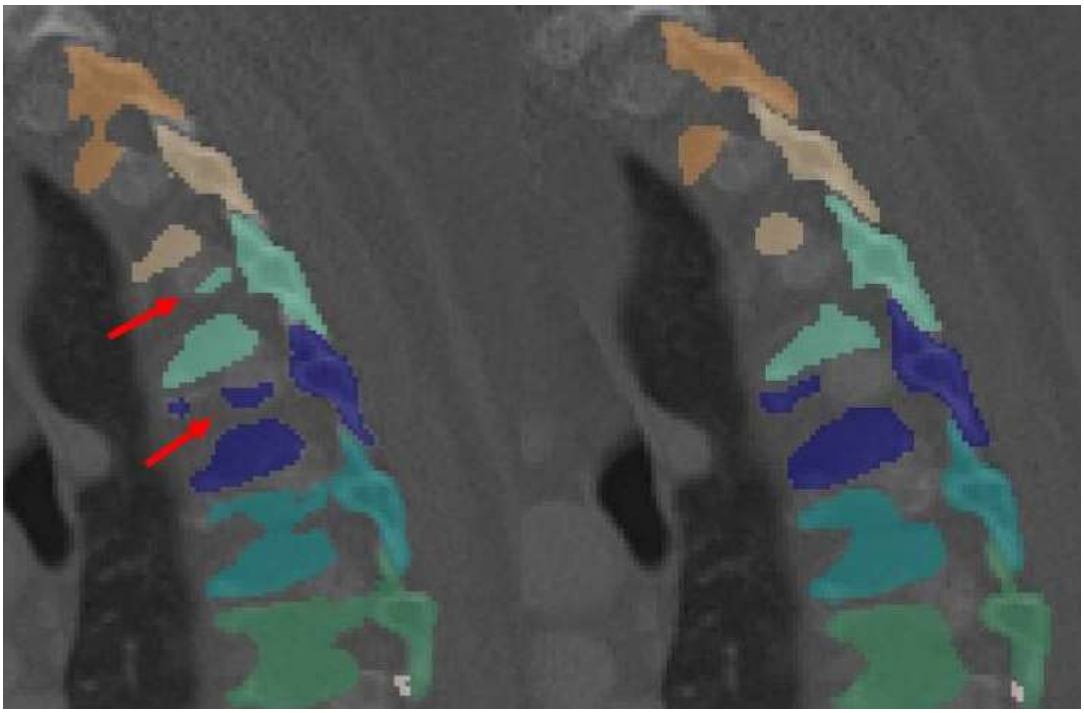

Abbildung 11: *Thorakaler Wirbelsäulenabschnitt in sagittaler Ansicht mit Segmentationsmaske des CT Verse ID13; Autosegmentierung, Pfeile demonstrieren Übersegmentierung am Rippenkopf (links), manuelle Korrektur der Segmentationsmaske am Rippenkopf (rechts)*

Die manuelle Segmentierung erfolgte für jeden WK in allen 3 Ebenen, d.h. sagittal, axial und koronal. Auf den Verzicht des 3D-Segmentierungsmodus der ITK- Snap- Software wurde explizit hingewiesen, da dieser zu einer unkontrollierbaren dreidimensionales Übersegmentierung nicht geforderter Areale führte. Das nachträgliche Eliminieren der

Übersegmentierung ist als Prozess wesentlich komplizierter und zeitaufwändiger als die Segmentierung an sich. Ein multiples Korrigieren in allen 3 Ebenen unter schwer abschätzbarer dreidimensionaler Auswirkung kann durch Ausschluss des 3D-Segmentierungsmodus verhindert werden. Im Bereich der Rippen ist eine Korrektur der Übersegmentierung unumgänglich. Die Separation der umgebenden ossären Strukturen zum WK zeigt eine erhöhte Fehleranfälligkeit des Algorithmus. Der nahtlose Übergang der Rippe mittels Rippengelenk zum Wirbelkörper forderte eine manuelle Differenzierung. (vgl. Abb. 12).

Ebenso bedarf die Unterscheidung des Facettengelenks sowohl zum angrenzenden WK als auch zu den Rippen in vielen Fällen einer nachträglichen Korrektur der Übersegmentierung. Die enge Lagebeziehung der Strukturen führt, ggf. auch unterstützt durch degenerative Veränderungen zu einem „Verschmelzen“ der Wirbelumrisse und geht mit einer ungenügenden computergestützten Segmentierungspräzision einher. (vgl. Abb.13)

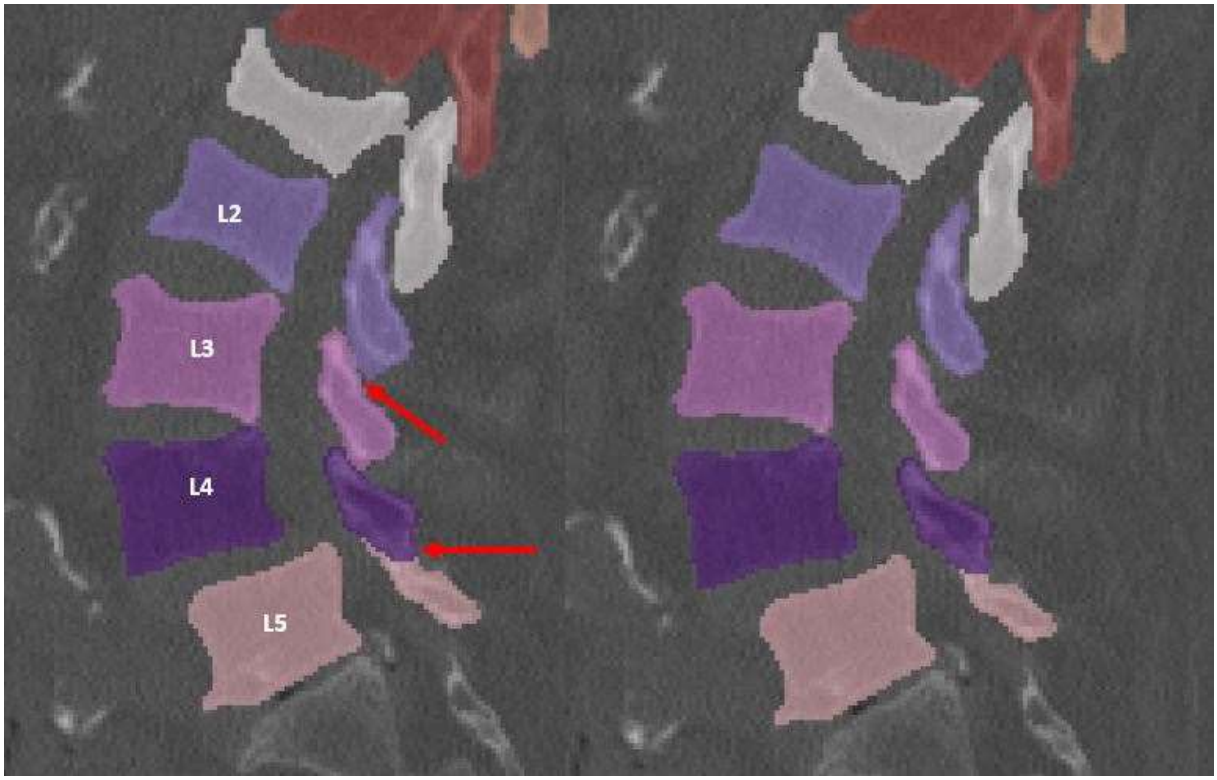

***Abbildung 12**: Thorakaler Wirbelsäulenabschnitt in sagittaler Ansicht mit Segmentationsmaske des CT Verse ID20 Computergenerierte Segmentationsmaske mit Übersegmentierung der Facettengelenke (links), manuelle Korrektur der Segmentationsmaske am Facettengelenk; Abgrenzung an L4 und L5 erschwert durch Verwachsung der Facettengelenke (rechts)*

Die Dice-Scores der Zweitkorrektur lassen somit rückschließen, dass nur kleine Korrekturen von JK getätigt wurden um ein präzises Segmentierungsergebnis zu generieren und zeigen, dass

nach Absprache wesentlicher Segmentierungskriterien eine geringe Variabilität zwischen den Ratern vorliegt. Da jedoch nur in vereinzelten Fällen ein Dice-Score von 1 für den Wirbelkörper erreicht wird, kann keine Rater-spezifische Reproduzierbarkeit generiert werden. Außerdem treten nach dem Experiment weitere grundsätzliche Fragestellungen auf. Wie verändert sich die Interrater-Variabilität in einem Versuch mit mehreren Korrektoren? Wird von allen Ratern ein vergleichbares Zeitpensum für die Korrektur gefordert? Kann die manuelle Segmentierung vom gleichen Rater reproduziert werden? Hat das Ausbildungsniveau des Rates Einfluss auf das Segmentierungssergebnis und welche Faktoren führen zu einer Rater-spezifischen Fehleranfälligkeit. Weitere Forschungsarbeiten sind hierfür notwendig um die Evidenz entsprechender Ergebnisse zu stärken.

4.4 Vergleich des Dice-Scores in Bezug auf die Wirbelsäulensegmente

In Kapitel 3.1.2. lässt sich anhand der Ergebnisse feststellen, dass die größte Fehleranfälligkeit des Algorithmus mit den Werten der Halswirbelsäule einhergeht. Dafür lassen sich mehrere Gründe anführen, welche die vergleichbar niedrigen Dice-Scores der HWS-Datensätze hervorrufen. Darunter zählen:

- Diskrepante Anatomie der Wirbelkörper im Vergleich mit der restlichen Wirbelsäule
- Enge Verwachsung zwischen Atlas und Axis
- Degenerative bzw. osteoporotische Veränderungen der Hals-Wirbelsäule mit morphologischer Variation der Wirbelkörper
- Fehlerhaft automatisch generierte Labels in der Segmentierungsmaske
- Rauschen, Artefakte und schlechte Bildqualität der Aufnahmen
- Geringere Anzahl an Trainingsdaten im Vergleich zur BWS & LWS

Durch die enge Lagebeziehung und das funktionelle Zusammenspielt zwischen Atlas und Axis (HW1 und HW2) hat die Abgrenzung der beiden Wirbelkörper einen erhöhten Schwierigkeitsgrad. Die im Durchschnitt niedrigeren Dice-Werte an HW1 und HW2 weisen auf eine verstärkte manuelle Nachkorrektur hin. Dabei zeigt sich, dass durch die anatomische Variation der beiden Wirbel eine vermehrt unpräzise Autosegmentierung resultiert. Der Algorithmus generiert exemplarisch in der Aufnahme DX033 eine lückenhafte Segmentierungsmaske. Die Unterscheidung der Wirbelkörper explizit an Axis und Atlas wird ungenügend umgesetzt und fälschlicherweise eine weitere Segmentierungsfarbe an HW2

eingefügt. (vgl. Abb. 13) In der Statistik äußert sich dies, durch das Auftreten eines zusätzlichen Wirbelkörpers mit vergleichbar stark erniedrigtem Dice- Score. Im Falle der DX033 erfolgt eine zusätzliche Auflistung eines WK10 (bei einer Segmentierung von lediglich 9 Wirbelkörpern) mit einem Dice- Score von 0,029. Daraus resultiert eine deutliche Verminderung des Mean-Dice-Scores, was unteranderem die Abweichungen zu den Datensätzen der BWS-LWS/ LWS Fälle mitbegründet. Fehlerhaft automatisch generierte Labels treten in fünf von insgesamt zwölf HWS-Datensätzen auf, sodass dies nicht als Einzelfall deklariert werden kann. Als Ursache dieser Problematik lässt sich ein fehlerhaftes Postprocessing des Algorithmus vermerken. Dabei indiziert die Überlagerung zweier Label an angrenzenden WK eine fehlerhafte Addition der Segmentierungsmasken und nachfolgend in der Produktion eines neuen Labels.

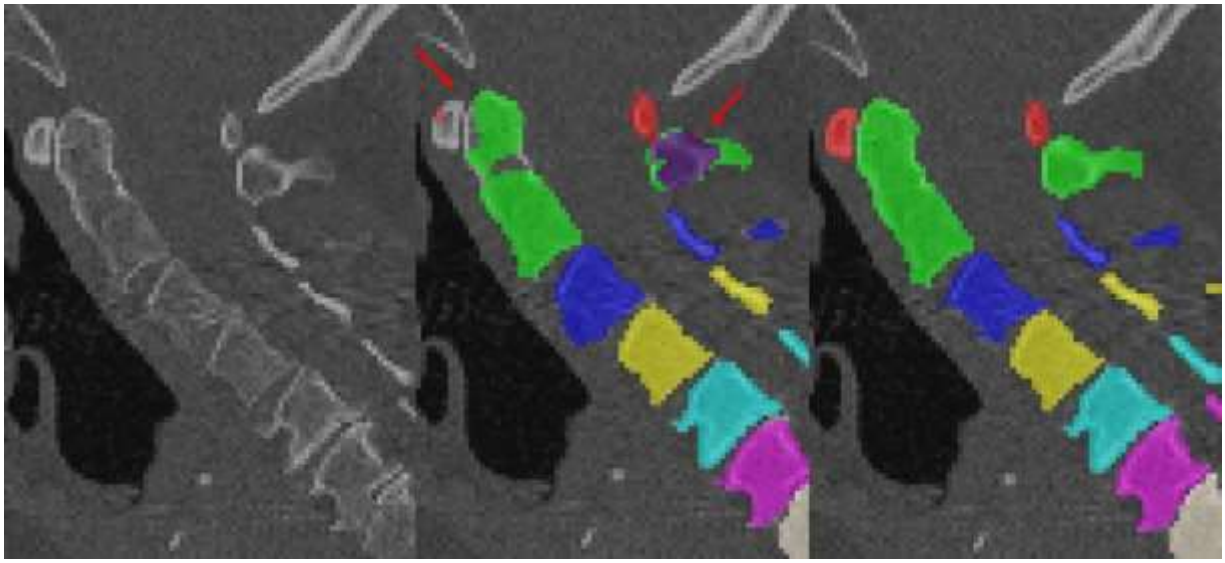

Abbildung 13*: Zervikaler Wirbelsäulenabschnitt in sagittaler Ansicht mit Segmentationsmaske des CT DX033; Darstellung des Ausgangs-CT (Links); Pfeile demonstrieren automatisierte unvollständige Segmentationsmaske an WK1 und WK2 mit zusätzlicher Segmentierungsfarbe am Dornfortsatz des WK2 (Mitte); Manuelle Korrektur der Segmentationsmaske an allen abgebildeten Wirbelkörpern und Extraktion der zusätzlichen Segmentierungsfarbe an WK2 (Rechts);*

Neben der natürlichen Formabweichung der Wirbel können jedoch auch degenerative Veränderungen das Erscheinungsbild der Wirbelsäule variieren. Dies lässt sich grundsätzlich für alle Wirbelkörperabschnitte beschreiben (vgl.4.2) und liefert in den betroffenen Datensätzen bei vorliegend ausgeprägten Verwachsungen, Degenerationen und osteoporotisch veränderten Wirbelkörpern erniedrigte Dice-Scores und Übersegmentierungen an den Wirbelgrenzen. Mit abnehmender Anzahl der Segmentierungen pro Wirbelsäulensegment steigt jedoch der Einfluss eines Patientenfalles auf den Meandice-Mittelwert der Wirbelsäulenfraktion. Dies folgert, dass

sich Dice-Score-Abweichungen im HWS-Datensatz mit 12 Segmentierungen verhältnismäßig stärker auswirken als im BWS-LWS-Datensatz mit 23 Segmentierungen. Die Evidenz des Vergleichs könnte durch eine einheitliche Anzahl der Segmentierungen pro Wirbelsäulenabschnitt verbessert werden.

Zuletzt soll auch der Einfluss von anatomisch unabhängigen Störfaktoren nicht außer Acht gelassen werden. Bild-Rauschen, niedriger Kontrast und Artefakte können sich negativ auf die Bildqualität der Aufnahmen auswirken. Dadurch besitzen sie einen nicht unerheblichen Einfluss auf die Genauigkeit der generierten Segmentierungsmaske. Die Präzision der Autosegmentierung und manuellen Segmentierung kann unter Verminderung der Bildqualität abnehmen. In Abb. 14 zeigt sich in Regio HWK3 ein deutlich ausgeprägtes Artefakt, welches mit einer Übersegmentierung des betroffenen Wirbelkörpers einhergeht. Grundsätzlich lässt sich dieser Aspekt auch für alle Wirbelsäulensegmente anführen.

Zusammenfassend konnten wir zeigen, dass eine automatische Segmentierung der Wirbelkörper derzeit nicht fehlerfrei umgesetzt werden kann. Die Autosegmentierung der HWS- Scans zeigte im Vergleich zur manuellen Erstkorrektur die größten Abweichungen. Die beste Leistung des CNN konnte in den Scans der Lendenwirbelsäule erzielt werden. Die Anzahl an vorhandenen Trainingsdaten korrelierend mit dem abgebildeten Wirbelsäulenabschnitt beeinflussen die Ergebnisse der Autosegmentierung. Zunehmende Variation des Bildmaterials und iterative Trainingseinheiten werden notwendig, um das algorithmische Outcome zu verbessern.

4.4.1 Einblick in weitere Segmentierungschallenges und Forschungsgebiete

Die von unserer Forschungsgruppe organisierte VerSe Challenge 2019 stellt eine von vielen derzeit durchgeführten Segmentierungschallenges dar. Weitere medizinische Fachrichtungen haben bereits öffentliche Datensets publiziert, um die Entwicklung von segmentierungsfähigen Algorithmen zu befördern. Zielführend soll die Forschung für KI und maschinelles Lernen weiter unterstützt werden. Um die Bandbreite an Anwendungsmöglichkeiten von CNN und das sich daraus ableitende diagnostische und therapeutische Potential zu untermauern, werden 4 weitere Forschungsgebiete angeführt.

Deep-Learning-Algorithmen können eine Segmentierung und Differenzierung von Arterien und Venen ermöglichen, um deren pathophysiologische Veränderungen zu untersuchen [35]. 2022 wurde die ASOCA - Challenge (Automated segmentation of normal and diseased coronary arteries) veranstaltet, um in koronarangiographischen CT- Aufnahmen mittels eines Algorithmus gesunde als auch pathologisch veränderte Koronararterien zu detektieren. Ziel des Wettbewerbs bestand in der Herstellung eines Benchmarktes für die Segmentierung von Koronararterien. Der Datensatz wurde öffentlich bereitgestellt, um zukünftig weitere Forschungsarbeiten zu ermöglichen und diagnostischer Maßnahmen langfristig zu simplifizieren.[36]

Ebenso finden Deep-Learning-Algorithmen zunehmend in der Tumorpathologie Anwendung. Tumordiagnostik, Subtypisierung, Staging und Grading sind Arbeitsschritte, welche unteranderem durch CNN's erleichtert werden sollen. [37] Vor diesem Hintergrund wurde in Zusammenarbeit mit den MICCAI- Konferenzen 2012 und 2013 der Brain-Tumor Image Segmentation Benchmark (BRATS) veranstaltet. Die Challenge forderte das Erstellen eines Algorithmus zur Segmentierung von Gliomen in einem öffentlich ausgeschriebenen Datenset mit 65 patientenspezifischen Multikontrast-MR-Scans. Die Leistungsfähigkeit von 20 eingereichten Algorithmen zur Visualisierung der Hirntumore wurde geprüft und anhand von Bewertungsmetriken der Gewinner mit dem besten Segmentierungsergebnis ermittelt. Eine Kombination der besten Algorithmen lieferte schlussendlich eine maßgebliche Verbesserung der Segmentierungsergebnisse. [38] Ebenfalls im Bereich der pathologischen künstlichen Intelligenz wurde 2019 durch die Pathology Artificial Intelligence Platform (PAIP) ein Wettbewerb organisiert. Ziel der Challenge bestand in der Konstruktion eines Algortihmus zur automatischen Detektion von Lebertumoren in Whole-Slide-Images (WSI). 28 Teilnehmer erfüllten diese Aufgaben fristgerecht. Dabei mussten neben der Visualisierung der karzinogenen Zellen das Ausmaß der lebensfähigen Tumorlast in 100 Datensätzen bestimmt werden. Die Veranstalter der Challenge erhoffen sich durch den Wettbewerb weitere Fortschritte im Bereich der virtuellen Diagnose von Lebertumoren. Ebenso soll langfristig die Wirksamkeit von onkologischen Therapieansätze in diesem Forschungsgebiet besser kontrolliert werden können. [39]

Als abschließendes Anwendungsbeispiel möchte ich auf das Projekt von M. Löffler et al. eingehen. Hier wurden mit der Unterstützung des in der Dissertation beschriebenen Algorithmus ein opportunistisches Screening-Tool (Anduin Bonescreen) erstmalig angewendet. Das Ziel der Forschungsarbeit lag in der Ausführung eines CNN zur automatischen Detektion von osteoporotischen Wirbelkörperfrakturen in CT Aufnahmen.

Diverse ossäre Parameter (Knochenmineralgehalt, trabelukäre und integral volumetrische Knochendichte etc.) wurden an segmentierten Wirbelkörpern automatisch und manuell bestimmt, um diese mit DXA- Messwerten zu vergleichen. Die Auswertung des Screening-Tools verzeichnete in Relation zu den vorliegenden DXA- Messungen eine bedeutend bessere Trennschärfe für die Prävalenz von osteoporotischen Frakturen. Die Autoren erwarten sich durch den Einsatz von Anduin-Bonescreen vielversprechende Ergebnisse, um zielführend die Diagnostik von osteoporotischen Knochenabbau und Wirbelkörperfrakturen entscheidend zu unterstützen. [40]

5 Zusammenfassung

In der Dissertation wurde ein dreiteiliges System zur Segmentierung der zervikalen, thorakalen und lumbalen Wirbelkörper in 160 CT-Scans beschrieben. Die Algorithmus-basierte Methode gliederte sich in Detektion, Labeling und Segmentierung der Wirbelkörper, woraus eine 3D-Multilabelmaske aller abgebildeter Wirbel resultierte. Das Ziel der Dissertation bestand darin die Leistungsfähigkeit eines computergestützten Verfahrens (Autosegmentierung) zu untersuchen und potentiell negative Einflüsse zu detektieren, welche eine manuelle Nachkorrektur der Segmentierung bedingten. Um die Präzision der Autosegmentierung zu beurteilen, wurde anhand 43 Ct-Scans jeweils ein Vergleich zu einer manuell korrigierten Segmentierungsmaske angestellt. Durch Erhebung des Dice-Scores pro Wirbelkörper konnte gezeigt werden, dass auf Grund der derzeit noch erhöhten Fehleranfälligkeit des Algorithmus erhebliche Abweichungen zwischen Autosegmentierung und Erstkorrektur vorlagen. Als potentielle Komplikation der Autosegmentierung wurden u.a. pathologische Veränderungen der Wirbel identifiziert. Osteophyten, Osteoporose und auch der Zustand nach Trauma mit ggf. vorliegender Spondylodese beeinträchtigten wesentlich den Outcome der Autosegmentierung. Jedoch auch die komplexe und individuelle Morphologie der Wirbelsäule wirkte sich nachteilig auf die Autosegmentierung aus. Diesbezüglich fand sich die größte Problematik an den Wirbelkörpern Atlas und Axis der Halswirbelsäule wieder. Die im Vergleich zur Brust-, und Lendenwirbelsäule deutlich erniedrigten Dice-Scores und das irrtümliche Einführen eines neuen Labels effizierten eine verminderte Segmentierungspräzision des zervikalen Wirbelsäulensegmentes. Auch die Qualität der medizinischen CT-Scans beeinflusste die Genauigkeit der Autosegmentierung. Bildspezifische Faktoren wie Artefakte, Rauschen und schlechte Kontrastierung haben einen deutlich negativen Einfluss auf die Segmentierung.

Die interindividuelle Korrektur der Segmentationsmasken mittels zweier Korrektoren konnte im Experiment eine insignifikante interindividuelle Abweichung der manuellen Segmentierung zeigen. Die konkrete Absprache relevanter anatomischer Strukturen, welche nach Autosegmentierung eine erhöhte manuelle Interaktion forderten, bedingte ein übereinstimmendes Segmentierungsergebnis. Darunter zählten Rippen und Facettengelenke in Abgrenzung zu den Wirbelkörpern. Auch der Verzicht des 3D-Segmentierungsmodus wirkte sich positiv auf die Segmentierungspräzision aus. Sofern in diesem Experiment nur zwei Korrektoren inkludiert wurden, wäre ein Vergleich mit multiplen Radiologen im Rahmen weiterführender Projekte wichtig, um die Interrater-Reliabilität zu verifizieren und

Rater-Spezifische Problematiken zu analysieren. Die manuelle Nachbearbeitung der Segmentierung forderte je nach Genauigkeit der Autosegmentierung, Komplexität der anatomischen Gegebenheit und pathologischen Variation der Wirbel einen erhöhten Zeitaufwand und eine intensive Auseinandersetzung mit dem Bildmaterial. Durch die progrediente Interkation mit der ITK-Snap Software und das Erlernen diverser Bearbeitungstools konnten schnellere Segmentierungsergebnisse erzielt werden. Allgemein lässt sich festhalten, dass mit zunehmender Anzahl an Trainingsdaten und Diversität des Bildmaterials der Algorithmus mittels Deep-Learning-Effekts die Autosegmentierung verbesserte.

Ausblick:

Derzeit verhindern unzureichende Trainingsdaten eine vollautomatische Segmentierung der Wirbelsäule und fordern eine humane Interaktion. Langfristig sollen jedoch bessere Algorithmen durch präzise und reproduzierbare Segmentierung der Wirbelkörper für diagnostische und therapeutische Zwecke genutzt werden. Die frühzeitige Detektion von krankhaft veränderten Wirbelkörpern, die prognostische Einschätzung von Wirbelfrakturen unter Osteoporose und auch die Überwachung von Alterungsprozessen im Sinne eines Staging sollen zukünftig durch jene computertechnischen Verfahren unterstützt und standardisiert werden.

6 Anhang

6.1 Dice- Koeffizient der Erstkorrektur

Dice-Koeffizient der Erstkorrektur

WK-Segment	Wirbelkörper	Mittelwert	Median	SD	Min	Max
HWS	1	0,865	0,911	0,125	0,497	0,964
	2	0,924	0,939	0,050	0,804	0,989
	3	0,935	0,958	0,059	0,770	0,984
	4	0,928	0,933	0,057	0,783	0,998
	5	0,935	0,946	0,047	0,802	0,994
	6	0,918	0,951	0,083	0,680	0,992
	7	0,943	0,962	0,069	0,683	0,996
BWS	8	0,865	0,911	0,125	0,497	0,964
	9	0,940	0,979	0,153	0,126	0,998
	10	0,838	0,962	0,317	0,000	1,000
	11	0,966	0,985	0,033	0,891	0,998
	12	0,970	0,987	0,028	0,918	0,998
	13	0,972	0,989	0,027	0,929	0,997
	14	0,974	0,988	0,025	0,924	0,999
	15	0,934	0,984	0,196	0,000	0,997
	16	0,977	0,989	0,022	0,932	0,999
	17	0,976	0,987	0,023	0,927	0,999
	18	0,968	0,987	0,043	0,794	0,998
	19	0,970	0,983	0,035	0,850	0,996
LWS	20	0,978	0,991	0,026	0,894	0,998
	21	0,976	0,993	0,035	0,828	0,999
	22	0,976	0,988	0,028	0,858	0,999
	23	0,978	0,994	0,023	0,929	0,999
	24	0,974	0,993	0,040	0,798	0,999
	25	0,998	0,998	0,000	0,998	0,998

6.2 Dice- Koeffizienten des Interrater

Dicekoeffizient des Interrater

WK-Segment	Wirbelkörper	Mittelwert	Median	SD	Min	Max
HWS	1	0,998	1,000	0,003	0,989	1,000
	2	0,999	1,000	0,002	0,995	1,000
	3	0,999	1,000	0,004	0,987	1,000
	4	0,999	1,000	0,002	0,996	1,000
	5	0,999	1,000	0,001	0,996	1,000
	6	0,997	1,000	0,008	0,970	1,000
	7	0,995	1,000	0,016	0,927	1,000
BWS	8	0,998	1,000	0,008	0,960	1,000
	9	1,000	1,000	0,001	0,998	1,000
	10	1,000	1,000	0,001	0,996	1,000
	11	0,999	1,000	0,002	0,992	1,000
	12	0,999	1,000	0,003	0,987	1,000
	13	0,999	1,000	0,002	0,988	1,000
	14	0,999	1,000	0,002	0,989	1,000
	15	0,999	1,000	0,002	0,990	1,000
	16	0,999	1,000	0,002	0,989	1,000
	17	0,999	1,000	0,003	0,985	1,000
	18	0,994	1,000	0,026	0,868	1,000
	19	0,998	1,000	0,004	0,985	1,000
LWS	20	1,000	1,000	0,000	0,999	1,000
	21	1,000	1,000	0,000	1,000	1,000
	22	1,000	1,000	0,001	0,998	1,000
	23	1,000	1,000	0,001	0,997	1,000
	24	0,998	1,000	0,005	0,979	1,000
	25	1,000	1,000	0,000	1,000	1,000

6.3 Für die Vergleiche herangezogene Patientendaten

Für die Vergleiche herangezogene Patientendaten				
Verse ID	**Seg. Pipeline**	**WS- Segment**	**ANZ. WK**	**Foreign. Material**
253	n8	BWS_LWS	17	0
257	n8	LWS	8	0
260	n8	BWS_LWS	16	0
261	n8	BWS_LWS	19	0
217	n8	HWS	9	0
264	n8	BWS_LWS	17	0
221	n8	HWS	9	0
265	n4aug	BWS_LWS	18	0
225	n9	HWS	9	0
226	n8	HWS	9	0
266	n8	BWS_LWS	17	0
227	n8	HWS	8	0
267	n8	BWS_LWS	17	0
230	n8	HWS	9	0
269	n8	BWS_LWS	16	0
232	n8	HWS	8	0
270	n8	BWS_LWS	18	0
235	n8	HWS	9	0
290	n8	BWS_LWS	19	0
271	n8	BWS_LWS	17	0
241	n8	HWS	9	0
273	n8	BWS_LWS	18	0
242	n8	HWS	8	0
276	n8	BWS_LWS	18	0
277	n8	BWS	16	0
278	n8	LWS	7	0
247	n8	HWS	8	0
279	n8	BWS_LWS	18	0
250	n8	HWS	9	0
205	n8	HWS	7	0
13	n9	BWS_LWS	17	0
16	n8	BWS_LWS	17	0
18	n8	BWS_LWS	16	0
22	n8	LWS	5	0
41	n8	BWS_LWS	18	0
64	n9	BWS_LWS	16	0
80	n9	BWS_LWS	15	0
93	n9	BWS_LWS	9	4
102	n9	LWS	5	0
146	n8	LWS	3	1
155	n8	BWS_LWS	15	0

6.4 Abbildungsverzeichnis

6.5 Tabellenverzeichnis

7 Literaturverzeichnis

1. Preisinger E. Diagnose Osteoporose. Man Medizin. 2014;52(3):214–20.

2. Gosch M, Stumpf U, Kammerlander C, Böcker W, Heppner HJ, Wicklein S. Management der Osteoporose nach Fragilitätsfrakturen. Z Gerontol Geriatr. 2018;51(1):113–25.

3. Madureira MM, Ciconelli RM, Pereira RMR. Quality of life measurements in patients with osteoporosis and fractures. Clinics. 2012;67(11):1315–20.

4. Mpotsaris A, Abdolvahabi R, Hoffleith B, Nickel J, Harati A, Loehr C, Gerdes CH, Hennigs S, Weber W. Perkutane Vertebroplastie von Wirbelkörperfrakturen benigner und maligner Genese: Eine prospektive Studie mit 1188 Patienten und einem Follow-Up von zwölf Monaten. Dtsch Arztebl. 2011;108(19):331–8.

5. Coughlan T, Dockery F. CME GERIATRIC MEDICINE Osteoporosis and fracture risk in older people. Clin Med (Northfield Il). 2014;14(2):187–91.

6. Chou SH, Vokes T. Vertebral Morphometry. J Clin Densitom [Internet]. 2016;19(1):S.48-53. Available from: http://dx.doi.org/10.1016/j.jocd.2015.08.005

7. Ross PD, Lyons A, Cooper C, Black D, Seeman E. Clinical consequences of vertebral fractures. Am J Med. 1997;103:S.30-43.

8. Bauer JS, Müller D, Rummeny EJ, Link TM. Frakturdiagnostik in der Osteoporose. Radiologe. 2006;46(10):S.839-846.

9. Raspe H, Raspe A, Holzmann M, Leidig G, Scheidt-Nave C, Felsenberg D, Banzer D, Matthis C. Die Reliabilität radiologischer Befunde zur Differentialdiagnose der vertebralen Osteoporose. Med Klin. 1998;93(SUPPL. 2):S.34-40.

10. Duff Putu, Jean Shoveller., Julio Montaner., Cindy Feng., Rachel Nicoletti., Kate Shannon . GO. 乳鼠心肌提取 HHS Public Access. Physiol Behav. 2016;176(1):139–48.

11. McCoy DB, Dupont SM, Gros C, Cohen-Adad J, Huie RJ, Ferguson A, Duong-Fernandez X, Thomas LH, Singh V, Narvid J, Pascual L, Kyritsis N, Beattie MS, Bresnahan JC, Dhall S, Whetstone W, Talbott JF. Convolutional neural network–based automated segmentation of the spinal cord and contusion injury: Deep learning

biomarker correlates of motor impairment in acute spinal cord injury. Am J Neuroradiol. 2019;40(4):737–44.

12. Zhou J, Damasceno PF, Chachad R, Cheung JR, Ballatori A, Lotz JC, Lazar AA, Link TM, Fields AJ, Krug R. Automatic vertebral body segmentation based on deep learning of dixon images for bone marrow fat fraction quantification. Front Endocrinol (Lausanne). 2020;11(September):1–10.

13. Kramme R. Medizintechnik: Verfahren Systeme Informationsverarbeitung [Internet]. Springer Berlin Heidelberg; 2013. S.596-598. Available from: https://books.google.de/books?id=I6LyBQAAQBAJ

14. Baumann T, Langer M. Bildnachverarbeitung Teil 1: Visualisierung und Segmentierung. Radiologe. 2013;53(9):805–9.

15. McGrath H, Li P, Dorent R, Bradford R, Saeed S, Bisdas S, Ourselin S, Shapey J, Vercauteren T. Manual segmentation versus semi-automated segmentation for quantifying vestibular schwannoma volume on MRI. Int J Comput Assist Radiol Surg [Internet]. 2020;15(9):1445–55. Available from: https://doi.org/10.1007/s11548-020-02222-y

16. Baumann T, Langer M. Bildnachverarbeitung Teil 2: Algorithmen und Workflow. Radiologe. 2013;53(12):1110–4.

17. Zaharchuk G, Gong E, Wintermark M, Rubin D, Langlotz CP. Deep learning in neuroradiology. Am J Neuroradiol. 2018;39(10):1776–84.

18. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Peter Campbell J. Introduction to machine learning, neural networks, and deep learning. Transl Vis Sci Technol. 2020;9(2):1–12.

19. Treder M, Eter N. Deep learning and neuronal networks in ophthalmology: Applications in the field of optical coherence tomography. Ophthalmologe. 2018;115(9):714–21.

20. Handels H. Medizinische Bildverarbeitung: Bildanalyse, Mustererkennung und Visualisierung für die computergestützte ärztliche Diagnostik und Therapie [Internet]. Vieweg+Teubner Verlag; 2009. S.95. (Leitfäden der Informatik). Available from: https://books.google.de/books?id=WdLYcz3rbT4C

21. Chan HP, Samala RK, Hadjiiski LM, Zhou C. Deep learning in medical image analysis. Adv Exp Med Biol. 2020;1213:3–21.

22. Peters KM. Osteoporose. 2013;S.403–415.

23. Kirschke JS, Sekuboyina A, Löffler MT. VerSe 2019 [Internet]. 2019. Available from: https://osf.io/nqjyw/

24. Sekuboyina A, Bayat A, Husseini ME, Löffler M, Li H, Tetteh G, Kukačka J, Payer C, Štern D, Urschler M, Chen M, Cheng D, Lessmann N, Hu Y, Wang T, Yang D, Xu D, Ambellan F, Amiranashvili T, Ehlke M, Lamecker H, Lehnert S, Lirio M, de Olaguer NP, Ramm H, Sahu M, Tack A, Zachow S, Jiang T, Ma X, Angerman C, Wang X, Wei Q, Brown K, Wolf M, Kirszenberg A, Puybareauq É, Valentinitsch A, Rempfler M, Menze BH, Kirschke JS. VerSe: A Vertebrae Labelling and Segmentation Benchmark for Multi-detector CT Images. 2020; Available from: http://arxiv.org/abs/2001.09193

25. Löffler MT, Sekuboyina A, Jacob A, Grau A-L, Scharr A, El Husseini M, Kallweit M, Zimmer C, Baum T, Kirschke JS. A Vertebral Segmentation Dataset with Fracture Grading. Radiol Artif Intell. 2020;2(4):e190138.

26. Sekuboyina A, Rempfler M, Valentinitsch A, Menze BH, Kirschke JS. Labeling Vertebrae with Two-dimensional Reformations of Multidetector CT Images: An Adversarial Approach for Incorporating Prior Knowledge of Spine Anatomy. Radiol Artif Intell. 2020;2(2):e190074.

27. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. CoRR; Med Image Comput Comput Interv { MICCAI 2015 [Internet]. 2015;abs/1505.0(Springer):234–41. Available from: http://arxiv.org/abs/1505.04597

28. Sekuboyina A, Rempfler M, Kukačka J, Tetteh G, Valentinitsch A, Kirschke JS, Menze BH. Btrfly Net: Vertebrae Labelling with Energy-Based Adversarial Learning of Local Spine Prior. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2018;11073 LNCS:649–57.

29. Hermsen M, Bel T, Boer M Den, Steenbergen EJ, Kers J, Florquin S, Roelofs JJTH, Stegall MD, Alexander MP, Smith BH, Smeets B, Hilbrands LB, Laak JAWMV. Deep learning-based histopathologic assessment of kidney tissue. J Am Soc Nephrol. 2019;30(10):1968–79.

30. Cohen J. A Power Primer Psychol Bull 112:155-159. Psychol Bull [PsycARTICLES] [Internet]. 1992;112(July):155–9. Available from: http://www2.psych.ubc.ca/~schaller/528Readings/Cohen1992.pdf

31. Mobbs RJ, Loganathan A, Yeung V, Rao PJ. Indications for anterior lumbar interbody fusion. Orthop Surg. 2013;5(3):153–63.

32. Kotsenas AL, Michalak GJ, DeLone DR, Diehn FE, Grant K, Halaweish AF, Krauss A, Raupach R, Schmidt B, McCollough CH, Fletcher JG. CT metal artifact reduction in the spine: Can an iterative reconstruction technique improve visualization? Am J Neuroradiol. 2015;36(11):2184–90.

33. van der Kraan PM, van den Berg WB. Osteophytes: relevance and biology. Osteoarthr Cartil. 2007;15(3):237–44.

34. Ezra D, Hershkovitz I, Salame K, Alperovitch-Najenson D, Slon V. Osteophytes in the Cervical Vertebral Bodies (C3–C7)—Demographical Perspectives. Anat Rec. 2019;302(2):226–31.

35. Nardelli P, Jimenez-Carretero D, Bermejo-Pelaez D, Washko GR, Rahaghi FN, Ledesma-Carbayo MJ, San Jose Estepar R. Pulmonary artery-vein classification in CT images using deep learning. IEEE Trans Med Imaging. 2018 Nov;37(11):2428–40.

36. Gharleghi R, Adikari D, Ellenberger K, Ooi SY, Ellis C, Chen CM, Gao R, He Y, Hussain R, Lee CY, Li J, Ma J, Nie Z, Oliveira B, Qi Y, Skandarani Y, Vilaça JL, Wang X, Yang S, Sowmya A, Beier S. Automated segmentation of normal and diseased coronary arteries – The ASOCA challenge. Comput Med Imaging Graph. 2022;97(February).

37. Jiang Y, Yang M, Wang S, Li X, Sun Y. Emerging role of deep learning-based artificial intelligence in tumor pathology. Cancer Commun. 2020;40(4):154–66.

38. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, Burren Y, Porz N, Slotboom J, Wiest R, Lanczi L, Gerstner E, Weber MA, Arbel T, Avants BB, Ayache N, Buendia P, Collins DL, Cordier N, Corso JJ, Criminisi A, Das T, Delingette H, Demiralp Ç, Durst CR, Dojat M, Doyle S, Festa J, Forbes F, Geremia E, Glocker B, Golland P, Guo X, Hamamci A, Iftekharuddin KM, Jena R, John NM, Konukoglu E, Lashkari D, Mariz JA, Meier R, Pereira S, Precup D, Price SJ, Raviv TR, Reza SMS, Ryan M, Sarikaya D, Schwartz L, Shin HC, Shotton J, Silva CA, Sousa N, Subbanna

NK, Szekely G, Taylor TJ, Thomas OM, Tustison NJ, Unal G, Vasseur F, Wintermark M, Ye DH, Zhao L, Zhao B, Zikic D, Prastawa M, Reyes M, Van Leemput K. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Trans Med Imaging. 2015;34(10):1993–2024.

39. Kim YJ, Jang H, Lee K, Park S, Min SG, Hong C, Park JH, Lee K, Kim J, Hong W, Jung H, Liu Y, Rajkumar H, Khened M, Krishnamurthi G, Yang S, Wang X, Han CH, Kwak JT, Ma J, Tang Z, Marami B, Zeineh J, Zhao Z, Heng PA, Schmitz R, Madesta F, Rösch T, Werner R, Tian J, Puybareau E, Bovio M, Zhang X, Zhu Y, Chun SY, Jeong WK, Park P, Choi J. PAIP 2019: Liver cancer segmentation challenge. Med Image Anal [Internet]. 2021;67:101854. Available from: https://doi.org/10.1016/j.media.2020.101854

40. Löffler MT, Jacob A, Scharr A, Sollmann N, Burian E, El Husseini M, Sekuboyina A, Tetteh G, Zimmer C, Gempt J, Baum T, Kirschke JS. Automatic opportunistic osteoporosis screening in routine CT: improved prediction of patients with prevalent vertebral fractures compared to DXA. Eur Radiol. 2021;31(8):6069–77.

8 Publikationen

A Vertebral Segmentation Dataset with Fracture Grading

Maximilian T Löffler [1], Anjany Sekuboyina [1], Alina Jacob [1], Anna-Lena Grau [1], Andreas Scharr [1], Malek El Husseini [1], Mareike Kallweit [1], Claus Zimmer [1], Thomas Baum [1], Jan S Kirschke [1]

Von der Abteilung für Diagnostische und Interventionelle Neuroradiologie, Medizinische Fakultät, Klinikum rechts der Isar, Technische Universität München Ismaninger Str. 22, München 81675, Germany (M.T.L., A. Sekuboyina, A.J., A.L.G., A. Scharr, M.E.H., M.K., C.Z., T.B., J.S.K.) & Lehrstuhl für Informatik, Technische Universität München, Deutschland (A. Sekuboyina)

Radiology Artificial Intelligence

Radiol Artif Intell. 2020 Jul; 2(4): e190138.

Veröffentlicht online 29.07.2020

doi: 10.1148/ryai.2020190138

VERSE: A Vertebrae labelling and segmentation benchmark for multi-detector CT images

Anjany Sekuboyina, Malek E. Husseini, Amirhossein Bayat, Maximilian Löffler, Hans Liebl, Hongwei Li, Giles Tetteh, Jan Kukačka, Christian Payer, Darko Štern, Martin Urschler, Maodong Chen, Dalong Cheng, Nikolas Lessmann, Yujin Hu, Tianfu Wang, Dong Yang, Daguang Xu, Felix Ambellan, Tamaz Amiranashvili, Moritz Ehlke, Hans Lamecker, Sebastian Lehnert, Marilia Lirio, Nicolás Pérez de Olaguer, Heiko Ramm, Manish Sahu, Alexander Tack, Stefan Zachow, Tao Jiang, Xinjun Ma, Christoph Angerman, Xin Wang, Kevin Brown, Alexandre Kirszenberg, Élodie Puybareau, Di Chen, Yiwei Bai, Brandon H. Rapazzo, Timyoas Yeah, Amber Zhang, Shangliang Xu, Feng Hou, Zhiqiang He, Chan Zeng, Zheng Xiangshang, Xu Liming, Tucker J. Netherton, Raymond P. Mumme, Laurence E. Court, Zixun Huang, Chenhang He, Li-Wen Wang, Sai Ho Ling, Lê Duy Huỳnh, Nicolas Boutry, Roman Jakubicek, Jiri Chmelik, Supriti Mulay, Mohanasankar Sivaprakasam, Johannes C. Paetzold, Suprosanna Shit, Ivan Ezhov, Benedikt Wiestler, Ben Glocker, Alexander Valentinitsch, Markus Rempfler, Björn H. Menze, Jan S. Kirschke

Medical Image Analysis, 2021 Oct.; Volume 73:102166

ISSN 1361-8415

doi: 10.1016/j.media.2021.102166

9 Danksagung

Zu guter Letzt ist es mir ein großes Vergnügen allen Personen zu danken, die mir bei der Fertigstellung meiner Doktorarbeit beigestanden haben. Nur durch deren Unterstützung ist mir die Fertigstellung meiner Doktorarbeit gelungen.

Besonderen Dank widme ich meinem Doktorvater Prof. Dr. med. Jan Kirschke, der mir diese Doktorarbeit anvertraut hat und zu jeder Zeit mit Muße, Hilfsbereitschaft und umfassender Betreuung zur Seite stand. Bei jeglicher Problematik wurde mir ein offenes Ohr geboten und diese mit Hilfe seinerseits stets zuverlässig und geduldig gelöst.

Des Weiteren möchte ich mich auch bei Anjany Sekuboyina und Bayat Amirhossein aus der Forschungsgruppe des Instituts für Radiologie – Abteilung für diagnostische und interventionelle Neuroradiologie des Klinikums rechts der Isar bedanken, welche mich bei der Erhebung meiner Forschungsdaten unterstützt haben. Durch ihr computertechnisches Knowhow konnte der Beitrag an der Verse-Challenge gestaltet und für die Doktorarbeit relevanten Daten zur Verfügung gestellt werden.

Zuletzt äußere ich ein herzliches Dankeschön an meine Familie und meinen Ehemann. Diese haben mir nicht nur kontinuierlich Unterstützung für die Doktorarbeit geboten, sondern bestärken mich bedingungslos in jeglichen Lebenslagen.